IntuiBeat:

Formative und summative Evaluation intuitiver Benutzung

Inaugural-Dissertation

zur Erlangung der Doktorwürde
der Fakultät für Humanwissenschaften der
Julius-Maximilians-Universität Würzburg

Vorgelegt von
Daniel Reinhardt aus Würzburg

Würzburg
2020

Erstgutachter:	Professor Dr. Jörn Hurtienne
Zweitgutachterin:	Professorin Dr. Carolin Wienrich

Tag des Kolloquiums: 11.09.2020

Danksagung

An dieser Stelle möchte ich gerne den Menschen danken, die mich während meiner Zeit als Doktorrand am Lehrstuhl für Psychologische Ergonomie an der Universität Würzburg unterstützt haben.

Als erstes möchte ich mich bei dir, **Jörn**, meinem Doktorvater und Mentor, bedanken. Du hast es geschafft, mit deiner unermüdlichen Leidenschaft einen Informatiker für die HCI und speziell für das Forschungsfeld der intuitiven Benutzung zu begeistern. Danke, dass du mir so viele Freiheiten in der Gestaltung meiner Forschung gelassen, mit mir gemeinsam verrückte Forschungsideen an unterschiedlichen Orten (z.B. Zweiviertel) entwickelt, und mir immer im richtigen Moment die passenden Impulse gegeben hast. Ich möchte auch dir, **Carolin**, meinen Dank für die Übernahme der Zweitbetreuung aussprechen. Durch deine fürsorgliche und lockere Art hat bei uns die Chemie sofort gestimmt. Danke für unseren erfolgreichen fachlichen Austausch und unsere tiefgründigen Gespräche über das Leben und seine Herausforderungen. Ich möchte mich außerdem noch speziell bei dir, **Tobi**, bedanken. Du hast es problemlos hinbekommen, mich als Student für eine wissenschaftliche Karriere zu begeistern und mir mit deiner Lässigkeit das Lampenfieber vor meinem ersten Konferenzvortrag zu nehmen. Dafür bin ich dir sehr dankbar.

Die im Rahmen dieser Dissertation durchgeführten Experimente wären nicht möglich gewesen ohne die Bemühungen einer Reihe weiterer einzigartiger Persönlichkeiten. Ich möchte mich herzlich bei meinen Studierenden **Jeremias**, **Benedict**, **Fabio**, **Sarah**, **Susanne** und **Miriam** für die Unterstützung bei der Datenerhebung bedanken. An dieser Stelle danke ich besonders dir, **Jeremias**. Danke für deine Anregungen und die Idee, die im Rahmen dieser Arbeit vorgestellte Methode schlicht IntuiBeat zu nennen. Ich möchte bei dieser Gelegenheit zudem noch den besten Kollegen der Welt **Sara**, **Franzi**, **Jan**, **Stephan**, **Nina**, **Clara** und **Cordula** danken. Jeder von euch hat auf seineihre eigene Art und Weise zum Gelingen dieser Arbeit beigetragen, egal ob durch einen künstlerischen Rat, ein produktives Gespräch oder eine lustige Unterhaltung bei einem Feierabendgetränk. Ich danke dir, **Frank**, für unsere netten Geek-Talks und die Unterstützung bei der Umsetzung jeglicher Hardware. Des Weiteren möchte ich dir, **Sandra**, dafür danken, dass du dich so gut um alle administrativen Belange kümmerst und ich dadurch mehr Zeit für diese Arbeit hatte. Ich danke außerdem dem Bundesministerium für Wirtschaft und Energie für die finanzielle Unterstützung im Zuge des Projekts 3D-GUIde und dem ganzen Projektteam. Hier bedanke ich mich besonders bei dir, **Michael**. Danke, dass du mir gezeigt hast, worauf es bei einer qualitativen Auswertung eines Usability-Tests wirklich ankommt und warum Norwegen ein so schönes Land ist. Ich bedanke mich an dieser Stelle auch bei dir, **Kristin**. Ohne dich, meiner Partnerin in Crime, wäre 3D-GUIde nicht dasselbe gewesen und ich vermisse unsere, fast täglichen, Zoom-Gespräche.

Nicht zuletzt möchte ich mich bei meinen Freunden für ihre Geduld, Ermutigungen und Zusprüche in den vergangenen Jahren bedanken. Ihr seid die Besten! Ich danke dir, **Priska**, für deine aufmunternden, liebevollen Worte während meiner Schreibphase und unsere vielen lustigen Ablenkungen. **Svenja**, ich danke dir für unsere

jahrelange Freundschaft, den Halt und deine unermüdliche Unterstützung dabei, dieser Arbeit den nötigen sprachlichen Glanz zu verleihen. Ich möchte auch dir, **Malin**, meinen herzlichsten Dank aussprechen. Danke, dass du auch in den letzten Tagen mutig genug warst, mir trotz meiner vielen Schachtelsätze wertvolles Feedback auf diese Arbeit zu geben. Ich bedanke mich für all die schönen Dinge, die wir gemeinsam erleben durften und wie du mich aufgrund deines einzigartigen Humors bewegst. **Linda**, bei dir bedanke ich mich dafür, dass du es irgendwie in jedem Moment geschafft hast, mir nicht das zu geben was ich gerade will, sondern das was ich in diesem Moment brauche. Ohne dich wären die letzten Wochen weniger einzigartig gewesen.

Abschließend möchte ich noch die Gelegenheit nutzen, mich bei meinen Geschwistern (**Julia** und **Marcus**) und meinen Eltern (**Marianne** und **Willi**) für ihre tiefe Liebe, die gemeinsamen Abenteuer, das unendliche Verständnis und die unzähligen Ratschläge im Laufe meines Lebens zu bedanken. Mama, we've made it!

Zusammenfassung

Intuitive Benutzung wird in dieser Arbeit definiert als das Ausmaß, mit dem ein Produkt mental effizient und effektiv genutzt wird, was mit einem starken metakognitiven Gefühl von Flüssigkeit einhergeht. Aktuelle Methoden verfügen nicht über eine ausreichend hohe zeitliche Anwendungseffizienz, um im Industrieprojekt 3D-GUIde effektiv zur Evaluation von Interaktionspatterns für 3D-Creation-Oriented-User-Interfaces (3D-CUIs) eingesetzt werden zu können. Diese Interaktionspatterns beschreiben strukturiert, wie 3D-CUIs als User Interfaces zur Erstellung von dreidimensionalen Inhalten gestaltet werden müssen, um intuitive Benutzung zu unterstützen. In dieser Arbeit werden daher zwei neue Evaluationsmethoden vorgeschlagen: 1) IntuiBeat-F als formative Evaluationsmethode und 2) IntuiBeat-S als summative Evaluationsmethode.

Basierend auf Default-Interventionist-Theorien und bestehenden Definitionen intuitiver Benutzung werden die mentale Beanspruchung als zentrales objektives, das metakognitive Gefühl von Flüssigkeit als zentrales subjektives und die Effektivität als zentrales pragmatisches mit intuitiver Benutzung assoziiertes Merkmal identifiziert. Die Evaluation intuitiver Benutzung mithilfe von IntuiBeat-F und IntuiBeat-S ist vielversprechend, da es sich bei beiden Methoden um Inhibition basierende Rhythmuszweitaufgaben handelt und diese somit mentale Beanspruchung objektiv erfassen können. Das Potential beider Methoden wird im Hinblick auf vorherige Forschungsarbeiten zur zeitlich effizienten Evaluation von 3D-CUIs aus der Mensch-Computer-Interaktion und der Psychologie diskutiert. Aus dieser Diskussion werden empirische Forschungsfragen abgeleitet.

Die erste Forschungsfrage untersucht die wissenschaftliche Güte von IntuiBeat-S. Im ersten, zweiten und dritten Experiment werden Paare von 3D-CUIs miteinander summativ verglichen (d.h. weniger vs. stärker intuitiv benutzbare User Interfaces). Dabei wird die wissenschaftliche Güte von IntuiBeat-S hinsichtlich der Hauptgütekriterien Objektivität, Reliabilität und Validität beurteilt. Die Ergebnisse zeigen, dass IntuiBeat-S eine hohe wissenschaftliche Güte bei der summativen Evaluation besitzt. Zudem macht es bei der Anwendung von IntuiBeat-S keinen Unterschied, ob der Rhythmus über die Ferse oder den Fußballen eingeben wird, und ob als Stichproben Studierende mit höherer oder geringerer Vorerfahrung bezüglich der Nutzung von 3D-CUIs verwendet werden.

Die zweite Forschungsfrage untersucht die wissenschaftliche Güte von IntuiBeat-F. Im vierten, fünften, sechsten und siebten Experiment werden 3D-CUIs einzeln formativ evaluiert (d.h. entweder ein weniger oder stärker intuitiv benutzbares User Interface). Dabei wird die wissenschaftliche Güte von IntuiBeat-F hinsichtlich der Hauptgütekriterien Gründlichkeit, Gültigkeit und Zuverlässigkeit beurteilt. Die Ergebnisse zeigen, dass IntuiBeat-F eine hohe wissenschaftliche Güte bei der formativen Evaluation besitzt. Diese liegt bei strikter Anwendung der Methode (d.h. Berücksichtigung ausschließlich mit der Methode entdeckter Nutzungsprobleme) zwar höher, ist aber bei wenig strikter Anwendung der Methode (d.h. Berücksichtigung auch unabhängig von der Methode entdeckter Nutzungsprobleme) noch ausreichend hoch. Jedoch konnte erst die Entwicklung und Einführung einer zusätzlichen Analysesoftware im Zuge des sechsten und siebten Experiments die wissenschaftliche Güte von IntuiBeat-F hinsichtlich aller drei Hauptgütekriterien demonstrieren, da ohne deren Unterstützung IntuiBeat-F vom Evaluator nicht ausreichend gründlich angewendet wird.

Die dritte Forschungsfrage untersucht, wie hoch die zeitliche Anwendungseffizienz beider Methoden als wichtiger Aspekt praktischer Güte im Vergleich zu bereits vorhandenen Evaluationsmethoden für intuitive Benutzung ist. Bezüglich der summativen Evaluation wird im zweiten Experiment eine höhere zeitliche Anwendungseffizienz von IntuiBeat-S im Vergleich zum aktuellen summativen Benchmark, der CHAI-Methode, sowohl bei der Evaluation von weniger als auch bei der von stärker intuitiv benutzbaren 3D-CUIs demonstriert. Auch bezüglich der formativen Evaluation konnten die Ergebnisse der letzten vier Experimente zeigen, dass die zeitliche Anwendungseffizienz von IntuiBeat-F im Vergleich zum aktuellen formativen Benchmark, dem Nutzertest mit retrospektivem Think-Aloud-Protokoll, sowohl bei der Evaluation von weniger als auch stärker intuitiv benutzbaren 3D-CUIs höher liegt. Dieser Unterschied bleibt bestehend, egal ob eine zusätzliche Analysesoftware vom Evaluator verwendet wird oder nicht.

Als Ergebnis aller Experimente lässt sich feststellen, dass die wissenschaftliche Güte und die zeitliche Anwendungseffizienz beider Methoden zur Evaluation intuitiver Benutzung von 3D-CUIs mehr als zufriedenstellend beurteilt werden kann. Die Arbeit wird mit einer Diskussion des geleisteten Forschungsbeitrags geschlossen. Dabei werden Anregungen für künftige Forschung aus theoretischer (z.B. Berücksichtigung des Gefühls von Flüssigkeit bei der Evaluation), praktischer (z.B. Untersuchung der Anwendbarkeit beider Methoden in anderen Domänen) und methodischer (z.B. Beurteilung der praktischen Güte beider Methoden anhand anderer Kriterien) Perspektive gegeben.

Abstract

In this work, intuitive use is defined as the extent to which a product can be used in a mentally efficient and effective manner, which is accompanied by a strong metacognitive feeling of fluency. Currently available methods to measure intuitive use lack sufficiently high time efficiency to apply them effectively for evaluating interaction patterns for 3D-Creation-Oriented-User-Interfaces (3D-CUIs) in the context of the industrial project 3D-GUIde. These interaction patterns describe in a structured way how 3D-CUIs – user interfaces aimed to create three-dimensional content – should be designed to support intuitive use. This work therefore proposes two new evaluation methods for intuitive use: 1) IntuiBeat-F as a formative method and 2) IntuiBeat-S as a summative method.

Based on default-interventionist theories and current definitions of intuitive, this work identifies mental efficiency as the central objective, the metacognitive feeling of fluency as the central subjective and the effectiveness as the central pragmatic attribute associated with intuitive use. The evaluation of intuitive use with IntuiBeat-F and IntuiBeat-S is promising since both methods are inhibition-based rhythm secondary tasks, and can therefore assess mental efficiency objectively. Their potential for a time efficient evaluation of 3D-CUIs is discussed in the light of existing research in human-computer interaction and psychology and empirical research questions are derived from this discussion.

The first research question examines the scientific quality of IntuiBeat-S. In the first, the second and the third experiment pairs of 3D-CUIs are compared in a summative manner (i.e., less intuitive vs. more intuitive user interfaces) and the scientific quality of IntuiBeat-S is assessed with regard to the scientific criteria objectivity, reliability and validity. The results show that IntuiBeat-S has a high scientific quality. Moreover, it makes no difference whether the rhythm is entered via the heel or the ball of the foot and whether the tested undergraduate students are more or less experienced in dealing with 3D-CUIs.

The second research question examines the scientific quality of IntuiBeat-F. In the fourth, the fifth, the sixth and the seventh experiment 3D-CUIs are evaluated individually in a formative manner (i.e., either a less intuitive or more intuitive user interface) and the scientific quality of IntuiBeat-F is assessed with regard to the scientific criteria thoroughness, validity and reliability. The results show that IntuiBeat-F has a high scientific quality. The scientific quality of IntuiBeat-F is higher when strictly applied (i.e., considering usability problems solely found with the method), but it is already high when applied less strictly (i.e., considering further usability problems found regardless of the method). However, only the development and use of an additional analysis software in the sixth and the seventh experiment show the scientific quality of IntuiBeat-F as a formative evaluation method for intuitive use with regard to all three scientific criteria. Without the software support, IntuiBeat-F is not applied thoroughly by the evaluator.

The third research question examines the practical quality of both methods in terms of application time efficiency when compared to existing evaluation methods for intuitive use. Regarding the summative evaluation, the second experiment demonstrates a higher application time efficiency of IntuiBeat-S in comparison to the current summative benchmark, the CHAI method, when evaluating both less and more intuitive 3D-CUIs. With regard to the formative evaluation, the last four experiments show a higher application

time efficiency of IntuiBeat-F in comparison to the current formative benchmark, the user test with retrospective think-aloud protocol, when evaluating both less and more intuitive 3D-CUIs, irrespective of whether or not an additional analysis software is used.

As a result of these studies it can be concluded that the scientific quality and the application time efficiency of both methods for the evaluation of intuitive use in the context of 3D-CUIs can be considered satisfactory. This work closes with a discussion of the accomplished research contribution and suggestions for future research from a theoretical (e.g., consideration of the feeling of fluency during the evaluation), a practical (e.g., applicability of both methods in other domains) and a methodological perspective (e.g., assessment of the practical quality of both methods using other criteria).

Inhaltsverzeichnis

1 Einleitung **13**
- 1.1 Hintergrund und Motivation . . . 13
- 1.2 IntuiBeat zur Evaluation intuitiver Benutzung . . . 19
- 1.3 Zielsetzung und Aufgabenstellung . . . 20
- 1.4 Überblick über die Arbeit . . . 21

2 Intuitive Benutzung **25**
- 2.1 Merkmale intuitiver Benutzung . . . 25
  - 2.1.1 IUUI-Forschergruppe . . . 26
  - 2.1.2 QUT-Forschergruppe . . . 51
  - 2.1.3 INTUI-Forschergruppe . . . 54
  - 2.1.4 Merkmale aus Sicht aller HCI-Forschergruppen . . . 62
- 2.2 Definierende Merkmale intuitiver Benutzung . . . 64
  - 2.2.1 Zwei-Prozess-Theorien . . . 65
  - 2.2.2 Arbeitsdefinition intuitiver Benutzung . . . 80
- 2.3 Zusammenfassung . . . 85

3 Evaluation intuitiver Benutzung **87**
- 3.1 Menschzentrierte Gestaltung interaktiver Systeme . . . 87
- 3.2 Evaluation in der HCI . . . 89
  - 3.2.1 Durchführung einer Evaluation . . . 90
  - 3.2.2 Ziele der Evaluation . . . 90
  - 3.2.3 Gestaltend-formative Evaluation . . . 91
  - 3.2.4 Bilanzierend-summative Evaluation . . . 99
  - 3.2.5 Meta-Evaluation in der HCI . . . 99
- 3.3 Formale Gütekriterien der formativen Evaluation . . . 106
  - 3.3.1 Hauptgütekriterium Gründlichkeit . . . 106
  - 3.3.2 Hauptgütekriterium Gültigkeit . . . 107
  - 3.3.3 Hauptgütekriterium Zuverlässigkeit . . . 108
  - 3.3.4 Nebengütekriterien . . . 108
- 3.4 Formale Gütekriterien der summativen Evaluation . . . 110
  - 3.4.1 Hauptgütekriterium Objektivität . . . 110
  - 3.4.2 Hauptgütekriterium Reliabilität . . . 112
  - 3.4.3 Hauptgütekriterium Validität . . . 113
  - 3.4.4 Nebengütekriterien . . . 117
- 3.5 Güte von Methoden zur formativen Evaluation intuitiver Benutzung . . . 119
- 3.6 Güte von Methoden zur summativen Evaluation intuitiver Benutzung . . . 120
  - 3.6.1 Subjektive Methoden . . . 122
  - 3.6.2 Objektive Methoden . . . 136
- 3.7 Zusammenfassung . . . 173

4 Evaluation intuitiver Benutzung mit IntuiBeat 177
4.1 Summative Evaluation mit IntuiBeat-S . . . 177
4.1.1 Empirische Befunde zur Güte der summativen Evaluation . . . 177
4.1.2 Diskussion von Limitationen und Bewältigung durch IntuiBeat-S . . 185
4.2 Formative Evaluation mit IntuiBeat-F . . . 186
4.2.1 Empirische Befunde zur Güte der formativen Evaluation . . . 186
4.2.2 Diskussion von Limitationen und Bewältigung durch IntuiBeat-F . . 189
4.3 Konkretisierung der Forschungsfragen . . . 190
4.4 Experimentelles Vorgehen . . . 192

5 Güte von IntuiBeat-S für die summative Evaluation intuitiver Benutzung 195
5.1 Experiment 1 . . . 197
5.1.1 Überprüfung der Objektivität von IntuiBeat-S . . . 197
5.1.2 Überprüfung der Reliabilität von IntuiBeat-S . . . 198
5.1.3 Überprüfung der Validität von IntuiBeat-S . . . 198
5.1.4 Methode . . . 201
5.1.5 Ergebnisse . . . 225
5.1.6 Diskussion . . . 231
5.1.7 Schlussfolgerung . . . 234
5.2 Experiment 2 . . . 235
5.2.1 Überprüfung der Validität von IntuiBeat-S . . . 236
5.2.2 Überprüfung der zeitlichen Anwendungseffizienz von IntuiBeat-S . . 238
5.2.3 Methode . . . 238
5.2.4 Ergebnisse . . . 251
5.2.5 Diskussion . . . 257
5.2.6 Schlussfolgerung . . . 260
5.3 Experiment 3 . . . 261
5.3.1 Überprüfung der Objektivität von IntuiBeat-S . . . 262
5.3.2 Überprüfung der Reliabilität von IntuiBeat-S . . . 262
5.3.3 Überprüfung der Validität von IntuiBeat-S . . . 262
5.3.4 Methode . . . 262
5.3.5 Ergebnisse . . . 272
5.3.6 Diskussion . . . 277
5.3.7 Schlussfolgerung . . . 279
5.4 Allgemeine Diskussion und Zusammenfassung . . . 280

6 Güte von IntuiBeat-F für die formative Evaluation intuitiver Benutzung 287
6.1 Experiment 4 . . . 289
6.1.1 Überprüfung der Gründlichkeit und Gültigkeit von IntuiBeat-F . . . 290
6.1.2 Überprüfung der zeitlichen Anwendungseffizienz von IntuiBeat-F . . 291
6.1.3 Methode . . . 291
6.1.4 Ergebnisse . . . 311
6.1.5 Diskussion . . . 320
6.1.6 Schlussfolgerung . . . 325
6.2 Experiment 5 . . . 326
6.2.1 Überprüfung der Gründlichkeit und Gültigkeit von IntuiBeat-F . . . 327
6.2.2 Überprüfung der zeitlichen Anwendungseffizienz von IntuiBeat-F . . 327

6.2.3 Methode . . . 327
6.2.4 Ergebnisse . . . 331
6.2.5 Diskussion . . . 339
6.2.6 Schlussfolgerung . . . 345
6.3 Experiment 6 . . . 346
6.3.1 Überprüfung der Gründlichkeit und Gültigkeit . . . 346
6.3.2 Überprüfung der zeitlichen Anwendungseffizienz . . . 347
6.3.3 Methode . . . 347
6.3.4 Ergebnisse . . . 352
6.3.5 Diskussion . . . 360
6.3.6 Schlussfolgerung . . . 366
6.4 Experiment 7 . . . 367
6.4.1 Überprüfung der Gründlichkeit und Gültigkeit . . . 368
6.4.2 Überprüfung der zeitlichen Anwendungseffizienz . . . 368
6.4.3 Methode . . . 368
6.4.4 Ergebnisse . . . 372
6.4.5 Diskussion . . . 379
6.4.6 Schlussfolgerung . . . 384
6.5 Allgemeine Diskussion und Zusammenfassung . . . 385

7 Zusammenfassende Diskussion und Ausblick 397
7.1 Zielsetzung der Arbeit: Entwicklung einer neuen Methode zur Evaluation intuitiver Benutzung . . . 397
7.2 Übersichtsarbeiten . . . 399
7.2.1 Übersicht über Definitionen intuitiver Benutzung . . . 399
7.2.2 Übersicht über Evaluationsmethoden intuitiver Benutzung . . . 402
7.3 Theoretischer Forschungsbeitrag . . . 405
7.4 Methodische Forschungsbeiträge . . . 410
7.4.1 IntuiBeat-S zur summativen Evaluation intuitiver Benutzung . . . 410
7.4.2 IntuiBeat-F zur formativen Evaluation intuitiver Benutzung . . . 411
7.5 Forschungsbeitrag im Sinne eines Werkzeugs . . . 412
7.6 Empirische Forschungsbeiträge . . . 415
7.6.1 Meta-Evaluation von IntuiBeat-S . . . 415
7.6.2 Meta-Evaluation von IntuiBeat-F . . . 418
7.7 Schlussfolgerung und Ausblick . . . 421

Literatur 423

Anhang 468

1 Einleitung

1.1 Hintergrund und Motivation

Der Begriff *intuitive Benutzung* hat sich in den letzten 10 Jahren von einem Schlagwort, das aufgrund seiner durchweg positiven Konnotation gern in Marketingkampagnen aufgegriffen wurde, zu einem eigenen Forschungsfeld in der Mensch-Maschine-Interaktion entwickelt (Blackler, 2018; Blackler & Popovic, 2015; Hurtienne, 2011; Reinhardt, Kuge, & Hurtienne, 2018; Ullrich, 2014). Durch die fortschreitende Digitalisierung der Berufswelt und des tägliches Lebens müssen in der heutigen Zeit Menschen, die in einer Vielzahl unterschiedlicher Bereiche tätig sind und damit auch ein unterschiedliches Vorwissen im Umgang mit Technologie besitzen (Reinhardt et al., 2018), täglich mit einer großen Anzahl an technischen Produkten interagieren, deren funktionelle Komplexität stetig steigt (Lawry, 2012; Lawry, Popovic, Blackler, & Thompson, 2019). Aufgrund der großen Anzahl an neuen komplexen Technologien, die in das tägliche Leben des Nutzers Einzug halten, sind die verfügbare Motivation und die Zeit für eine Einarbeitung in die jeweilige Technologie trotz steigender Abhängigkeit stark reduziert (Hurtienne, 2011).

Die Gestaltung eines intuitiv benutzbaren Produkts, Dienstes oder Systems, das keine vorherige Einarbeitung benötigt, kann eine Möglichkeit darstellen, mit diesem technischen Wandel umzugehen und so eine integrative Gesellschaft zu fördern (Reinhardt et al., 2018). Intuitive Benutzung liegt dann vor, wenn sich die Technologie für den Nutzer vertraut anfühlt (Blackler, 2018; Raskin, 1994), der Nutzer dabei nicht über die Interaktion mit der Technologie nachdenken muss (Blackler, 2018; Reinhardt et al., 2018) und er die Technologie stattdessen auf seinem Bauchgefühl basierend benutzen kann, ohne die Handhabung der Technologie explizit gelernt zu haben (Mohs, Hurtienne, Kindsmüller, Israel, Meyer et al., 2006; Ullrich & Diefenbach, 2011). Typische Definitionen geben an, dass intuitive Benutzung einem Nutzer gestattet, mit einem technischen Produkt, einem Dienst, oder einem System effektiv und kognitiv mühelos zu interagieren, aufgrund der eigenen unbewussten Anwendung von Vorwissen (Blackler, 2008, 2018; Blackler & Popovic, 2015; Hurtienne, 2011; O'Brien, Rogers, & Fisk, 2010).

Der beschriebene technische Wandel lässt sich unter anderem auch im Bereich der 3D-User-Interfaces beobachten. Aufgrund des technologischen Fortschritts in den letzten 20 Jahren ist hierfür leistungsfähige und gleichzeitig erschwingliche Hardware (z.B. Computer, Smartphones, Tablets, diverse Ein- und Ausgabegeräte) entstanden, was der Entwicklung und Anwendung von 3D-User-Interfaces in verschiedenen Anwendungsdomänen (z.B. Unterhaltung, Fertigung, Erziehung, Kunst, Informationsvisualisierung, Simulation) neue Wege erschlossen hat. 3D-User-Interfaces ermöglichen Nutzern, die aus verschiedenen Bereichen stammen, in einer virtuellen dreidimensionalen Umgebung zu interagieren (Burmester et al., 2018; LaViola, Kruijff, McMahan, Bowman, & Poupyrev, 2017; Preim & Dachselt, 2015). Die Aufgaben, die ein Nutzer in 3D mithilfe diverser Eingabegeräte durchführt, sind vielfältig und umfassen dabei die Selektion von Objekten, die (geometrische)

Manipulation von Objekten (z.B. Objektplatzierung und Objektrotation), die Navigation in der virtuellen Umgebung (z.B. Veränderung der Ansicht bzw. der Kameraperspektive) und die Systemsteuerung (z.B. Texteingabe und Menüauswahl). Die mit diesen Aufgaben verbundenen *Interaktionstechniken* werden gemäß ihrer unterstützenden Aufgabe (d.h. Selektionstechniken, Manipulationstechniken, Navigationstechniken und Systemkontrolltechniken) benannt (LaViola et al., 2017; Preim & Dachselt, 2015).

Obwohl heute die technologische Entwicklung soweit fortgeschritten ist, dass Systeme eine 3D-Interaktion in Echtzeit mit aufwendigen 3D-Modellen unterstützen können, bleibt das Problem bestehen, wie Nutzer letztendlich Objekte in diesen dreidimensionalen virtuellen Umgebungen intuitiv benutzen können. Die Ursache dieses Problems liegt darin begründet, dass ein Nutzer implizit davon ausgeht, dass er mit den realistischen 3D-Modellen auf die gleiche Art und Weise interagieren kann, wie er es in der realen Welt tut (LaViola et al., 2017). Intuitive Benutzung wird im Rahmen dieser Arbeit definiert als das Ausmaß, mit dem ein Produkt mental effizient und effektiv genutzt wird, was mit einem starken metakognitiven Gefühl von Flüssigkeit einhergeht (siehe Teilabschnitt 2.2.2). Insbesondere die Forderung nach einer mental effizienten Interaktion kann in der heutigen Zeit von aktuellen 3D-User-Interfaces nicht vollständig erfüllt werden und eine mögliche Erklärung für Probleme bei der 3D-Interaktion liefern, da Nutzer immer noch Schwierigkeiten bei der Wahrnehmung, Verarbeitung und Interpretation von dreidimensionalen Inhalten haben. Zudem verstehen sie die damit verbundenen Interaktionen nicht, aufgrund mangelnder Hinweisreize in der virtuellen Umgebung, die ihnen in der Realität zur Verfügung stehen (LaViola et al., 2017).

Egal, ob (1) dem Nutzer die 3D-Interaktion in einem CAD-Programm (d.h. Computer-Aided Design; Programm zur Erstellung von 3D-Modellen) lediglich über einen einfachen LCD-Monitor mit Maus und Tastatur gestattet wird (siehe Hand, 1997), (2) sich der Nutzer über sein Smartphone oder Tablet computergenerierte, die Realität überlagernde, Zusatzinformationen zu einem Exponat im Museum anzeigen lässt (d.h. erweiterte Realität, siehe Azuma, 1997), oder (3) der Nutzer ein Head-Mounted Display zur stereoskopischen Darstellung eines virtuellen interaktiven Kreißsaals verwendet, wo er mithilfe zweier Controller die Durchführung einer Operation trainiert, die virtuelle Umgebung wirkt für den Nutzer zwar immer irgendwie dreidimensional, verliert aber durch die Darstellung auf zweidimensionalen Oberflächen und der indirekten Interaktion mit den virtuellen Inhalten über vorhandene Eingabegeräte einiges an wichtigen Informationen (z.B. unterschiedliche Tiefenwahrnehmung, fehlende physische Einschränkungen, fehlendes multisensorisches Feedback, siehe Hinckley, Pausch, Goble, & Kassell, 1994). Das erschwert die Übertragung des aus der echten Welt stammenden Vorwissens auf die virtuelle 3D-Umgebung, und beeinträchtigt so eine mental effiziente Benutzung (LaViola et al., 2017; Preim & Dachselt, 2015).

Da bereits etablierte 2D-Interaktionstechniken (z.B. Dropdown-Listenfelder) nur bedingt bei 3D-User-Interfaces anwendbar sind (siehe Hand, 1997; Jankowski & Hachet, 2015; Preim & Dachselt, 2015), müssen neue intuitive *Interaktionslösungen* entwickelt werden, um diese Lücke zu schließen und den Nutzern eine ganzheitliche intuitive Benutzung ermöglichen zu können (Burmester et al., 2018). 3D-User-Interfaces und damit verbundene Dienstleistungen werden in Deutschland von rund 2500 Unternehmen bereitgestellt, bei

denen es sich größtenteils um kleine und mittelständische Unternehmen (d.h. KMUs) handelt (Astor, Jarowinsky, & von Lukas, 2013). Jedoch konzentrieren sich diese hauptsächlich auf die technische Verbesserung der 3D-Darstellung und vernachlässigen dabei die Bereitstellung intuitiver Interaktionslösungen (Burmester et al., 2018).

Usability bzw. Gebrauchstauglichkeit wird innerhalb der DIN EN ISO 9241-11 definiert als „das Ausmaß, in dem ein Produkt durch bestimmte Benutzer in einem bestimmten Nutzungskontext genutzt werden kann, um bestimmte Ziele effektiv, effizient und zufriedenstellend zu erreichen“ (ISO, 1998, S. 4). Intuitive Benutzung bildet einen Teilaspekt davon, der sich überwiegend der mentalen Beanspruchung bei der Benutzung von Technologie widmet (Hurtienne, 2011). Usability und somit auch intuitive Benutzung wird von vielen KMUs in Deutschland oftmals als bedeutsamer Wettbewerbsfaktor vernachlässigt, und es werden dementsprechend wenig Ressourcen für die damit verbundene Forschung und Entwicklung bereitgestellt (Burmester et al., 2018; Woywode, Mädche, Wallach, & Plach, 2012). Wo sich im Bereich von 2D-User-Interfaces schon längst *Interaktionspatterns*, welche häufiges und bewährtes Gestaltungswissen strukturiert beschreiben (Alexander, 1977; Dearden & Finlay, 2006), durchgesetzt haben und dazu unzählige Interaktionspatternsammlungen existieren, auf deren Grundlage eine intuitive Gestaltung theoretisch möglich wäre (z.B. Borchers & Thomas, 2001; Crumlish & Malone, 2009; Scott & Neil, 2009; Tidwell, 2010; Van Duyne, Landay, & Hong, 2007; Van Welie & Van der Veer, 2003; Wesson & Cowley, 2003), fehlen für 3D-User-Interfaces aufgrund des stetigen technologischen Wandels und der generellen hohen Komplexität des Forschungsfelds solche Patternsammlungen (Burmester et al., 2018; Figueroa & Castro, 2011; M. Green & Lo, 2004). KMUs, die keine Kapazitäten für eine ganzheitliche Gestaltung und Evaluation ihrer User Interfaces aufbringen können, können somit nicht auf bewährtes Gestaltungswissen zurückgreifen, was sich wiederum negativ auf die intuitive Benutzung ihrer User Interfaces auswirkt (Burmester et al., 2018).

Im Rahmen des Projekts *3D-GUIde* (Laufzeit 2015 - 2018), welches in der Initiative „Einfach intuitiv - Usability für den Mittelstand“ im Förderschwerpunkt „Mittelstand-Digital“ des Bundesministeriums für Wirtschaft und Energie gefördert wurde, wurde deswegen versucht, eine Interaktionspatternsammlung für die einfachste Form von 3D-User-Interfaces zu erstellen, die herkömmliche Desktopsysteme (d.h. bestehend aus Monitor, Tastatur und Maus, siehe Hand, 1997) für die Interaktion nutzen. Im Forschungsprojekt wurden außerdem Interaktionspatterns für Virtual Reality und Augmented Reality entwickelt. Interaktionspatterns für Desktopsysteme bildeten jedoch das Herzstück des Projekts, da die auf Desktopsystemen häufig genutzten *Creation-Oriented-User-Interfaces* (CUIs) bei KMUs den höchsten Verbreitungsgrad in Deutschland besitzen (Astor et al., 2013) und solche KMUs durch die Förderinitiative speziell gefördert werden sollten (Burmester et al., 2018). Unter dem Begriff CUIs fallen laut Akers (2010) alle User Interfaces, deren zentrale Zielsetzung das Erstellen von Inhalten ist. Im Bereich von 3D-User-Interfaces fallen darunter alle Arten von 3D-Modellierungssoftware (d.h. Computer-Aided-Design: computergestütztes Konstruieren und Entwickeln von zwei- und dreidimensionalen Modellen), mit denen sich laut dem Bericht von Astor et al. (2013) immer noch die meisten 3D-Unternehmen in Deutschland in den Bereichen Architektur, Bauwesen, Maschinenbau und Anlagenbau beschäftigen. Andere Beispiele für CUIs aus dem 2D-Bereich wären Bildbearbeitungs- und Textverarbeitungsprogramme (Akers, 2010; Akers, Jeffries, Simpson, & Winograd, 2012).

Um für 3D-CUIs eine Interaktionspatternsammlung zu entwickeln, wurden im Projekt basierend auf den Erkenntnissen aus einer Nutzungskontextanalyse (siehe Beyer & Holtzblatt, 1997; Holtzblatt & Beyer, 2016) vielversprechende Patternkandidaten (d.h. Interaktionslösungen für diverse Aufgaben in verschiedenen verwendeten CUIs) durch Experteneinschätzung identifiziert (z.B. 3D-CUI-Interaktionslösungen für Rotieren und Verschieben von 3D-Objekten), welche im Nachgang auf ihre intuitive Benutzbarkeit evaluiert werden mussten (Burmester et al., 2018). Eine solche Evaluation bezüglich intuitiver Benutzung bildet im Gegensatz zu bereits existierenden Sammlungen aus dem 2D-Bereich eine Besonderheit, da die im Rahmen des Projekts entwickelte 3D-Interaktionspatternsammlung nicht nur theoretisch zur intuitiven Gestaltung befähigen sollte, sondern darin enthaltene Lösungen auch bereits empirisch bezüglich intuitiver Benutzung evaluiert sind (Burmester et al., 2018). Um derartiges leisten zu können, wurde im Projekt zum einen eine *formative Evaluationsmethode* benötigt, um damit Nutzungsprobleme bei der Systeminteraktion aufzudecken und so die getesteten Patternkandidaten entweder anhand des qualitativen Feedbacks zu verbessern oder sie begründet verwerfen zu können. Zum anderen wurde eine *summative Evaluationsmethode* für eine abschließende Bewertung der verbesserten Interaktionslösungen benötigt, um diese als Patterns in der Interaktionspatternsammlung festzuschreiben und die Verbesserung bezüglich intuitiver Benutzung quantitativ belegen zu können (Burmester et al., 2018).

Die im Rahmen des Projekts eingesetzten formativen und summativen Evaluationsmethoden müssen als Mindestvoraussetzung wissenschaftlich tragfähige Ergebnisse liefern, weswegen ihre *wissenschaftliche Güte* durch die Sicherstellung bestimmter formaler Hauptgütekriterien (z.B. formative wissenschaftliche Güte durch Gründlichkeit und summative wissenschaftliche Güte durch Validität, siehe Abschnitte 3.3 und 3.4) nachgewiesen sein muss, was durch einen Vergleich mit allgemeingültigen quantitativen Normen erfolgt (siehe Döring & Bortz, 2016). Aufgrund der Tatsache, dass im Rahmen des Projekts für die Entwicklung von Interaktionspatterns viele 3D-CUIs untersucht werden mussten (Burmester et al., 2018), wurde bei der Auswahl von formativen und summativen Evaluationsmethoden zusätzlich auf deren zeitliche Anwendungseffizienz (d.h. Zeit für Durchführung, Auswertung und Interpretation) geachtet, welche auch von Schmidt-Kretschmer und Blessing (2006) als eine essentielle Anforderung an eine die Gestaltung eines Systems unterstützende Methodik genannt wird. Die Anforderung an eine Methode, zeitlich effizient angewendet werden zu können, stellt ein Nebengütekriterium von formativen und summativen Evaluationsmethoden dar, welches Methoden erfüllen müssen, um im Anwenderprojekt praktisch eingesetzt werden zu können. Anhand dieses Nebengütekriteriums wird ein Teilaspekt der *praktischen Güte* sichergestellt, die den praktischen Nutzen der Methode aus der Sicht des Anwenders reflektiert. Zur Sicherstellung der praktischen Güte können neben der zeitlichen Anwendungseffizienz abhängig von der untersuchten Art der Evaluationsmethode (d.h. formativ oder summativ) noch weitere Aspekte genutzt werden, um diese vollständig erfassen zu können (z.B. formative praktische Güte durch Downstream Utility und summative praktische Güte durch Testfairness, siehe Abschnitte 3.3 und 3.4).

Der Hintergrund für diese Anforderung nach zeitlicher Anwendungseffizienz an eine formative bzw. summative Evaluationsmethode und die damit verbundene Fokussierung auf lediglich einen Teilaspekt praktischer Güte liegt an der hohen Diversität der Systemnutzung, die für CUIs charakteristisch ist. Das liegt daran, dass der Verwendungszweck für solche Arten von Software sehr allgemein gehalten und die inhaltlichen Möglichkeiten theo-

retisch unbeschränkt sind (Akers, 2010; Akers et al., 2012). Mit einem CAD-Programm wie AutoCAD können beispielsweise Maschinen modelliert, aber auch einfache zweidimensionale Vektorgrafiken erstellt werden. CUIs bieten für die Erledigung dieser Aufgaben diverse Lösungswege an, die mit verschiedenen Interaktionssequenzen (z.B. Shortcut zum Öffnen eines Untermenüs vs. sequentielles Auswählen des Untermenüs mithilfe mehrerer Mausklicks) und Interaktionstechniken (z.B. Verändern der Ansicht über ein Interaktionselement oder über eine Tastenkombination) assoziiert sind. Beispielsweise kann ein Nutzer bei der Modellierung eines 3D-Modells grundsätzlich eine eher additive (d.h. ein 3D-Objekt entsteht durch das Zusammenfügen einer oder mehrerer einfacher dreidimensionaler Grundformen) oder subtraktive Modellierungsstrategie (d.h. ein 3D-Objekt entsteht durch das Wegnehmen einer oder mehrerer einfacher dreidimensionaler Grundformen) verwenden (Gebhardt, 2017). Der Verlauf der Systemnutzung wird dementsprechend von den inhaltlichen Zielen und den gewählten Strategien des Nutzers bestimmt (Akers, 2010; Akers et al., 2012), wodurch die resultierende Systemnutzung bei verschiedenen Nutzern des Systems variieren kann. Deswegen wurde der zeitlichen Anwendungseffizienz bei der Evaluation dieser Interaktionen im Rahmen des Anwenderprojekts ein besonderer Stellenwert zugeordnet, um möglichst alle Facetten der Systemnutzung bei der Patternentwicklung berücksichtigen zu können.

Laut eines aktuellen Reviews von Blackler, Popovic und Desai (2018) kommen zur summativen Evaluation intuitiver Benutzung überwiegend subjektive Methoden (d.h. Fragebögen zur subjektiven Selbstbeurteilung) zum Einsatz. Es können aber nicht ausschließlich diese Methoden zur Evaluation von Gestaltungselementen und Gestaltungsaspekten eingesetzt werden, da sich diese retrospektiv, also erst nach der Systemnutzung beim Nutzer erfassen lassen und man deswegen mit hoher Sicherheit nur von Nutzer bewusst wahrgenommene Gestaltungselemente und Gestaltungsaspekte evaluieren kann. Kognitive Strategieänderungen bei der Aufteilung der mentalen Ressourcen während der eigentlichen Systemnutzung können damit nicht erkannt werden. Um zusätzlichen Zugriff auf die unbewusst wahrgenommenen Aspekte der Systeminteraktion zu haben, sollten zusätzlich objektive Methoden genutzt werden (siehe Abschnitt 3.6). Als objektive Methoden kommen im Forschungsfeld intuitiver Benutzung überwiegend Verhaltensmaße (d.h. Verhaltenseinschätzungen durch Experten) und Hauptaufgabenleistungsmaße (d.h. Maße der Effektivität wie der Anteil an korrekt gelösten Aufgaben) zum Einsatz (Blackler et al., 2018).

Bei Verhaltensmaßen wird das Nutzerverhalten durch Expertenbeurteilung objektiv bewertet bzw. vorausgesagt. Als Grundlage fungiert dabei häufig ein Nutzertest (Blackler et al., 2018). Bei einem Nutzertest wird das System mit repräsentativen Endanwendern anhand realistischer Aufgaben objektiv evaluiert, was ein quasi-experimentelles Verfahren, mit oder ohne eingeschränkte Bedingungsvariation darstellt (Sarodnick & Brau, 2006). Die Expertenbeurteilung ist dabei üblicherweise sehr zeitaufwendig, da der Experte hierfür eine Reihe von Parametern (d.h. Beurteilungsschemata) bewerten muss, die ihm erlauben, das Ausmaß an intuitiver Benutzung abzuschätzen (Reinhardt & Hurtienne, 2017; Reinhardt et al., 2018). Hauptaufgabenleistungsmaße leiden wie subjektive Methoden ebenfalls daran, kognitive Strategieänderungen bei der Aufteilung der mentalen Ressourcen während der eigentlichen Systeminteraktion nicht direkt abbilden zu können. Aufgrund der Tatsache, dass Menschen eine erhöhte mentale Aufgabenbelastung durch eine größere Anstrengung kompensieren können (G. F. Wilson, 2005), ist die Beurteilung intuitiver Benutzung durch Hauptaufgabenleistungsmaße alleine nicht ausreichend. Insbesondere in

Situationen, in denen der Nutzer nur einer geringen mentalen Belastung ausgesetzt ist und diese dementsprechend durch vermehrte Anstrengung kompensieren kann, ist bei der Interpretation von Effektivität Vorsicht geboten (Cain, 2007; F. Chen et al., 2016; Hockey, 1997; Wierwille & Eggemeier, 1993).

Da sich intuitive Benutzung anhand der mentalen Beanspruchung, also der mentalen Effizienz bei der Systemnutzung objektiv bewerten lässt, können ergänzend objektive Methoden zum Einsatz kommen, die zwar schon außerhalb der Forschung zu intuitiver Benutzung in der HCI (d.h. Human-Computer Interaction; Mensch-Maschine-Interaktion) bei der Evaluation mentaler Beanspruchung angewendet wurden, aber noch nicht explizit im Forschungsbereich zu intuitiver Benutzung zum Einsatz kamen. Hierbei sind besonders physiologische Maße und Zweitaufgabenleistungsmaße zu nennen (Reinhardt & Hurtienne, 2017). Vielversprechende physiologische Maße, wie die Herzratenvariabilität, können anfällig für motorische Artefakte und den generellen Erregungszustand des Nutzers sein (Charles & Nixon, 2019). Dies führt wiederum zu einem schlechten Signal-Rausch-Verhältnis, weswegen Messungen mithilfe aufwendiger Algorithmen zuvor auf Kosten der zeitlichen Anwendungseffizienz der Methode bereinigt werden müssen (Charles & Nixon, 2019). Zweitaufgabenleistungsmaße haben damit zu kämpfen, dass durch das Einführen einer zweiten, während der eigentlichen Aufgabe ausgeführten Nebenaufgabe, die Art und Weise beeinflusst werden kann, wie der Nutzer seine Hauptaufgabe (d.h. seine eigentliche Systemnutzung) erledigt (z.B. kann ein Nutzer bei seiner Hauptaufgabe durch eine Zweitaufgabe unterbrochen oder abgelenkt werden). Um Zweitaufgaben weniger intrusiv zu machen, müssen Haupt- und Zweitaufgabe genau aufeinander abgestimmt werden, was ebenfalls einen hohen zeitlichen Aufwand bedeutet und sich negativ auf die zeitliche Anwendungseffizienz der Methode auswirken kann (Reinhardt & Hurtienne, 2017).

Laut dem bereits erwähnten aktuellen Review von Blackler et al. (2018) kommen zur formativen Evaluation lediglich die Methode des Think-Aloud-Protokolls (d.h. „lautes Denken“) in verschiedenen Ausführungen bei einem Nutzertest zum Einsatz (siehe Abschnitt 3.5). Bei einem Think-Aloud-Protokoll wird der Nutzer aufgefordert, seine Gedankengänge und Handlungsstrategien bei der Systemnutzung laut zu verbalisieren. Ein Experte kann auf diese Weise Probleme bei der Systemnutzung anhand der Offenlegung der kognitiven Informationsverarbeitung des Nutzers erkennen. Es kann dabei grundsätzlich zwischen einem parallelen und einem retrospektiven Protokoll unterschieden werden (Nielsen, 1994). Das parallele Ausführen eines Think-Aloud-Protokolls während der Aufgabenbearbeitung wird dabei am häufigsten als formative Evaluationsmethode eingesetzt (Blackler et al., 2018). Jedoch kann der Einsatz eines parallelen Think-Aloud-Protokolls die wissenschaftliche Güte der Methode gefährden, da durch das parallele Ausführen des Protokolls eine Zweitaufgabe eingeführt wird, die zwar bestenfalls die Einsicht fördern kann, aber schlimmstenfalls die Ausführung der Hauptaufgabe (d.h. Systemnutzung) negativ beeinflussen (z.B. unterbrechen) und somit ebenfalls einen negativen Einfluss auf die gefundenen Probleme haben kann. Dementsprechend sollte immer auf ein retrospektives Protokoll zurückgegriffen werden (Reinhardt & Hurtienne, 2018), in dem ein Nutzer retrospektiv eine Videoaufzeichnung seines Nutzertests kommentiert und so dem Experten beim Aufspüren von Problemen hilft. Dieses Vorgehen beeinträchtigt jedoch die zeitliche Effizienz bei der Anwendung der Methode, weswegen diese Art des Think-Aloud-Protokolls auch kaum bei der formativen Evaluation zum Einsatz kommt (siehe Blackler et al., 2018).

Zusammenfassend lässt sich für aktuell verfügbare formative und summative Evaluationsmethoden im Forschungsbereich zu intuitiver Benutzung festhalten, dass diese entweder aufgrund ihrer schlechten zeitlichen Anwendungseffizienz oder ihrer methodischen Einschränkungen, die die Güte der Methode gefährden können, nicht für die Evaluation von User Interfaces (speziell 3D-CUI-Interaktionslösungen im Rahmen des Projekts 3D-GUIde) geeignet sind. Aufgrund dieser Problematik lässt sich für das Projekt 3D-GUIde ein Bedarf erkennen, der sich in der zentralen Forschungsfrage dieser Dissertation widerspiegelt: Wie kann die intuitive Benutzung von User Interfaces (speziell 3D-CUIs) sowohl formativ als auch summativ mit möglichst hoher zeitlicher Anwendungseffizienz evaluiert werden?

1.2 IntuiBeat zur Evaluation intuitiver Benutzung

Im Rahmen dieser Dissertation wird daher die neue objektive Evaluationsmethode *IntuiBeat* vorgestellt, mit der eine formative (d.h. IntuiBeat-F) und summative (d.h. IntuiBeat-S) Evaluation intuitiver Benutzung auf zeitlich effizientere Weise als mit bisherigen Ansätzen möglich ist. Um die intuitive Benutzung bei der Nutzung eines User Interfaces (d.h. Nutzung eines CUI als Hauptaufgabe) kontinuierlich beurteilen zu können, muss der Nutzer einen zuvor erlernten Rhythmus als Zweitaufgabe während seiner eigentlichen Hauptaufgabe mit seinem Fuß klopfen.

Eine Reihe von empirischen Arbeiten konnte die wissenschaftliche Güte von Rhythmusabweichungen (d.h. Abweichung zwischen tatsächlich geklopftem Rhythmus während der Durchführung der Hauptaufgabe und einer zuvor erhobenen individuellen Baseline ohne parallele Hauptaufgabe) bereits als summative Evaluationsmethode für mentale Beanspruchung im E-Learning-Bereich formal bezüglich der genannten summativen Hauptgütekriterien nachweisen (z.B. Korbach, Brünken, & Park, 2018; Park & Brünken, 2015). Es konnte jedoch noch nicht explizit demonstriert werden, ob Rhythmusabweichungen, speziell im Bereich der Mensch-Maschine-Interaktion, für die summative Evaluation intuitiver Benutzung von User Interfaces (speziell 3D-CUIs) mit ausreichender wissenschaftlicher Güte genutzt werden können (siehe Abschnitt 4.1). Zusammenfassend lässt sich deswegen festhalten, dass die wissenschaftliche Güte von Rhythmusabweichungen als summative Evaluationsmethode *IntuiBeat-S* für intuitive Benutzung im Rahmen dieser Dissertation sichergestellt werden muss. Ferner sollen dabei die speziellen Anforderungen von CUIs berücksichtigt werden, um IntuiBeat-S als eine summative Evaluationsmethode im Projekt 3D-GUIde einsetzen und ihre zeitliche Anwendungseffizienz gegenüber bisherigen, im Forschungsfeld der intuitiven Benutzung verfügbaren, Ansätzen demonstrieren zu können.

Auch als formative Evaluationsmethode *IntuiBeat-F* lassen sich Rhythmusabweichungen nicht bedenkenlos einsetzen, obwohl es naheliegend ist, dass sich damit auch Probleme bei der Systemnutzung, also die Ursachen für eine nicht intuitive Benutzung, erkennen lassen. Es sind hierzu einige empirische Befunde aus der Forschung zu mentaler Beanspruchung im Bereich der Mensch-Maschine-Interaktion vorhanden (z.B. M. J. Albers, 2011; Tracy & Albers, 2006). Diese Befunde konnten aufgrund einer Reihe von methodischen Einschränkungen jedoch die wissenschaftliche Güte von Rhythmusabweichungen nicht formal bezüglich der genannten formativen Hauptgütekriterien nachweisen. Es konnte auf diese Weise dementsprechend nicht explizit gezeigt werden, ob Rhythmusabweichungen für die

formative Evaluation intuitiver Benutzung von User Interfaces (speziell 3D-CUIs) mit ausreichender wissenschaftlicher Güte genutzt werden können (siehe Abschnitt 4.2). Zusammenfassend lässt sich deswegen festhalten, dass die wissenschaftliche Güte von Rhythmusabweichungen als formative Evaluationsmethode *IntuiBeat-F* für intuitive Benutzung im Rahmen dieser Dissertation sichergestellt werden muss. Ferner sollen dabei die speziellen Anforderungen von CUIs berücksichtigt werden, um IntuiBeat-F als eine formative Evaluationsmethode im Projekt 3D-GUIde einsetzen und ihre zeitliche Anwendungseffizienz gegenüber bisherigen, im Forschungsfeld zu intuitiver Benutzung verfügbaren, Ansätzen demonstrieren zu können.

1.3 Zielsetzung und Aufgabenstellung

Die Zielsetzung dieser Dissertation besteht daher in der Entwicklung und Evaluation einer formativ und summativ im Projekt 3D-GUIde einsetzbaren objektiven Evaluationsmethode für intuitive Benutzung namens IntuiBeat, deren

- wissenschaftliche Güte als summative Evaluationsmethode *IntuiBeat-S* für intuitive Benutzung nachgewiesen ist.
- wissenschaftliche Güte als formative Evaluationsmethode *IntuiBeat-F* für intuitive Benutzung nachgewiesen ist.
- zeitliche Anwendungseffizienz als formative und summative Evaluationsmethode für intuitive Benutzung und damit ein insbesondere im Projekt 3D-GUIde, wichtiger Teilaspekt praktischer Güte nachgewiesen ist.

Im Zuge dessen müssen zuvor jedoch die folgenden Bedingungen erfüllt sein:

- Erarbeiten einer den aktuellen Forschungsstand zusammenfassenden Arbeitsdefinition von intuitiver Benutzung (d.h. Messdefinition) durch Prüfen verfügbarer HCI-Definitionen und weiterführender Arbeiten, um theoretisch zu fundieren, dass sich intuitive Benutzung als das Ausmaß, mit dem ein Produkt mental effizient und effektiv genutzt wird, was mit einem starken metakognitiven Gefühl von Flüssigkeit einhergeht, überhaupt abbilden und damit evaluieren lässt.
- Entwicklung von IntuiBeat als formative (d.h. IntuiBeat-F) und summative Evaluationsmethode für intuitive Benutzung (d.h. IntuiBeat-S) basierend auf bisherigen theoretischen und empirischen Erkenntnissen zur Abbildung intuitiver Benutzung durch Rhythmusabweichungen, sowie Bereitstellung geeigneter Soft- und Hardware zur Aufzeichnung und Analyse, um auf diese Weise eine objektive und methodisch korrekte Messung garantieren zu können.
- Zur Überprüfung der wissenschaftlichen und praktischen Güte von IntuiBeat als formative (d.h. IntuiBeat-F) und summative Evaluationsmethode (d.h. IntuiBeat-S), müssen sowohl ein geeigneter formaler Evaluationsprozess (d.h. formale Meta-Evaluation) als auch formale Gütekriterien und für diese Kriterien geeignete Außenkriterien anhand von vorhandenen empirischen Befunden aus der Literatur abgeleitet werden.

Nachdem die theoretische Grundlage für diese Arbeit durch die Erfüllung dieser Vorbedingungen gelegt wurde, konnten sieben empirische Experimente durchgeführt werden, um die wissenschaftliche und zeitliche Anwendungseffizienz von IntuiBeat sicherzustellen. Nachdem dieses Ziel erfüllt wurde, wurde mithilfe von IntuiBeat eine Reihe von Patternkandidaten im Zuge von 3D-GUIde formativ und summativ evaluiert. Auf Basis der Ergebnisse konnten 3D-Interaktionspatterns festgeschrieben und die fertige 3D-Interaktionspatternsammlung interessiertene KMUs über eine Webseite (http://3d-guide.net) bereitgestellt werden. Dieser Aspekt wird jedoch im Rahmen dieser Arbeit nicht mehr thematisiert, da das Projekt lediglich als Rahmenbedingung für diese Arbeit fungiert und nicht den wissenschaftlichen Beitrag, die Entwicklung und Meta-Evaluation der Methode *IntuiBeat*, darstellt. Für eine genaue Beschreibung des Vorgehens im Anwenderprojekt und der dort erarbeiteten Patternsammlung wird daher auf Burmester et al. (2018) verwiesen.

1.4 Überblick über die Arbeit

Kapitel 2 dient der vertieften Betrachtung der Konstrukte *mentale Beanspruchung* und *intuitive Benutzung*, wobei ersteres, wie bereits erwähnt, das zentrale Merkmal darstellt, mit dem intuitive Benutzung objektiv bewertet werden kann. Einen Schwerpunkt des zweiten Kapitels bildet das Aufzeigen verfügbarer Definitionen intuitiver Benutzung aus dem HCI-Bereich, der Diskussion deren empirischer Basis und die metatheoretische Verknüpfung dieser Definitionen mithilfe der Handlungsregulationstheorie. Basierend auf diesen Definitionen und der metatheoretischen Betrachtung werden Merkmale zusammengetragen, die typischerweise mit intuitiver Benutzung im Bereich der Mensch-Maschine-Interaktion in Verbindung gebracht werden. Auf Basis aktueller Forschung aus dem Bereich der Zwei-Prozess-Theorien werden anschließend die drei zentralen definierenden Merkmale identifiziert, mit der zuverlässig das Ausmaß an intuitiver Benutzung aus einer objektiven, subjektiven und pragmatischen Perspektive bewertet werden kann. Auf Grundlage des dadurch identifizierten subjektiven zentralen Merkmals (d.h. metakognitives Gefühl von Flüssigkeit), des objektiven zentralen Merkmals (d.h. mentale Beanspruchung) und des ergänzenden pragmatischen Merkmals (d.h. Effektivität) wird abschließend eine Arbeitsdefinition intuitiver Benutzung aus einer Handlungsregulationsperspektive für diese Arbeit bereitgestellt, auf deren Basis intuitive Benutzung anhand von IntuiBeat formativ und summativ evaluiert werden kann. Es handelt sich bei dieser Definition daher um eine Messdefinition.

In Kapitel 3 wird zunächst der Begriff der Evaluation eingeführt und aufgezeigt, wo eine Evaluation im Rahmen eines menschenzentrierten Gestaltungsprozesses verortet werden kann. Dabei wird hervorgehoben, dass Evaluationsmethoden grundsätzlich eine qualitative bzw. formative oder eine summative bzw. quantitative Ausrichtung besitzen können, die auch deren Einsatzzeitpunkt innerhalb eines menschzentrierten Gestaltungsprozesses bestimmt. Abhängig davon, ob es sich um eine formative oder eine summative Evaluation handelt, müssen bei der Evaluation einer neuen Methode bzw. der Bewertung einer bereits verfügbaren Methode unterschiedliche formale Gütekriterien berücksichtigt werden, um valide Aussagen bezüglich der wissenschaftlichen und praktischen Güte (z.B. zeitliche Anwendungseffizienz) der Methode unter Berücksichtigung von Evaluationszielen und

des Anwenderinteresses machen zu können (d.h. Durchführung einer Meta-Evaluation). In der Literatur vorgeschlagene formale Gütekriterien werden nachfolgend für formative und summative Evaluationsmethoden vorgestellt und eine Auswahl auf Basis der Anforderungen aus dem Projekt 3D-GUIde getroffen. Abschließend wird der aktuelle Forschungsstand bezüglich formativer und summativer Evaluationsmethoden präsentiert und unter Berücksichtigung der Anforderungen aus dem Projekt 3D-GUIde bewertet (d.h. Sicherstellung wissenschaftlicher Güte und zeitlicher Anwendungseffizienz). Ferner werden unter diesem Hintergrund Limitationen gesammelt, die die im Zentrum dieser Arbeit stehende Entwicklung und Meta-Evaluation der neuen Methode *IntuiBeat* aus einer formativen (d.h. IntuiBeat-F) und summativen Perspektive (d.h. IntuiBeat-S) rechtfertigen können. Im Zuge der Bewertung des Forschungsstands aus Anwendungsprojektsicht werden auch geeignete formative und summative Evaluationsmethoden als Außenkriterien für die Evaluation von IntuiBeat vorgestellt, sowie die aktuellen formativen und summativen Benchmarks des Feldes identifiziert. Als Benchmarks sind objektive etablierte Methoden zu verstehen, deren wissenschaftliche Güte anderen im Feld verfügbaren Methoden überlegen ist.

Kapitel 4 beschreibt wie Rhythmusabweichungen zur formativen und summativen Evaluation intuitiver Benutzung genutzt werden können, und in welchem Zusammenhang sie zu intuitiver Benutzung stehen. Das Kapitel betrachtet dazu zunächst die theoretische Grundlage für diesen angenommenen Zusammenhang und stellt dazu bereits vorhandene empirische Belege für die wissenschaftliche Güte von Rhythmusabweichungen als Indikator für intuitive Benutzung aus der Forschung zu mentaler Beanspruchung vor. Dabei werden die Forschungslücken aufgezeigt, die die Zielsetzung und Aufgabenstellung dieser Arbeit darstellen. Das Kapitel endet mit einer Zusammenfassung der in den vorherigen Kapiteln diskutierten Inhalte und einer Konkretisierung der Forschungsfragen dieser Arbeit. Im Zuge dieser Konkretisierung wird auch das empirische Vorgehen zu deren Beantwortung aufgezeigt.

Die nächsten beiden Kapitel umfassen die Darstellung der empirischen Arbeiten, die sich thematisch in zwei Teile untergliedern lassen. Im ersten empirischen Teil, welcher in Kapitel 5 beschrieben wird, wird die wissenschaftliche Güte von IntuiBeat-S als summative Evaluationsmethode anhand der in Kapitel 3 definierten formalen Gütekriterien mit ausgewählten summativen Außenkriterien bei diversen CUIs/Aufgaben demonstriert (Experiment 1 bis 3). Dabei fungiert die Expertenbeurteilung mithilfe eines Beurteilungsschemas (d.h. CHAI-Methode, siehe Reinhardt et al., 2018) als das objektive Benchmark-Kriterium der summativen Evaluation. Es werden bei der Meta-Evaluation aber auch zusätzlich Ergebnisse weiterer subjektiver und objektiver Maße als Außenkriterien berücksichtigt, die in Kapitel 3 ebenfalls vorgestellt wurden. In Experiment 2 erfolgt zusätzlich die Sicherstellung der zeitlichen Anwendungseffizienz durch den Vergleich von IntuiBeat-S mit dem objektiven Benchmark-Kriterium. Kapitel 5 endet damit, dass Ergebnisse bezüglich der wissenschaftlichen Güte und zeitlichen Anwendungseffizienz diskutiert, sowie methodische Probleme erörtert und Anregungen für die künftige Forschung bezüglich IntuiBeat-S gegeben werden.

Im zweiten empirischen Teil, welcher in Kapitel 6 beschrieben wird, wird die wissenschaftliche Güte von IntuiBeat-F als formative Evaluationsmethode anhand der in Kapitel 3 definierten formalen Gütekriterien mit einem ausgewählten formativen Außenkriterium

bei diversen CUIs/Aufgaben demonstriert (Experiment 4 bis 7). Dabei fungiert ein klassischer Nutzertest mit retrospektivem Think-Aloud-Protokoll als das objektive Benchmark-Kriterium und als alleiniges formatives Außenkriterium (siehe Blackler et al., 2018). An dieser Stelle kann nur auf ein Außenkriterium zurückgegriffen werden, da die formative Evaluation im Forschungsfeld zu intuitiver Benutzung bislang noch nicht stark thematisiert wurde. Deswegen stellt dieses Außenkriterium auch gleichzeitig den Benchmark des Feldes dar. In allen vier Experimenten erfolgt zusätzlich auch ein Vergleich von IntuiBeat-F und dem objektiven Benchmark-Kriterium bezüglich zeitlicher Anwendungseffizienz. Kapitel 6 endet wie das Kapitel zuvor mit einer Diskussion der Ergebnisse bezüglich der wissenschaftlichen Güte und zeitlichen Anwendungseffizienz von IntuiBeat-F, sowie damit verbundener methodischer Probleme. Es werden abschließend Anregungen für die künftige Forschung bezüglich IntuiBeat-F gegeben.

In der zusammenfassenden Diskussion in Kapitel 7 werden IntuiBeat-F als formative und IntuiBeat-S als summative Evaluationsmethode für intuitive Benutzung bei CUIs kritisch reflektiert, methodische Probleme thematisiert, Implikationen für die Evaluation intuitiver Benutzung außerhalb von CUIs diskutiert und Vorschläge für künftige Forschungsarbeiten aufgezeigt. An dieser Stelle werden auch die durch diese Arbeit geleisteten Forschungsbeiträge für den Bereich der Mensch-Maschine-Interaktion erläutert. Schließlich werden im Anhang studienübergreifende (z.B. Rekrutierung der Versuchspersonen, Teilnehmerinformation) und studienspezifische Zusatzinformationen (z.B. Aufgaben, Fragebögen) bereitgestellt.

Abschließend soll noch erwähnt werden, dass die im Zuge der gesamten Arbeit gewählte männliche Form immer gleichermaßen weibliche oder diverse Personen mit einbezieht. Es wurde daher auf eine konsequente Doppelbezeichnung aufgrund besserer Lesbarkeit verzichtet.

2 Intuitive Benutzung

Intuitiver Benutzung wird ein immer größerer Stellenwert in der HCI zugeordnet, um heterogenen Nutzergruppen aufgrund der fortschreitenden Digitalisierung trotz ihres unterschiedlichen Vorwissens eine „einfache“ Interaktion mit neuen Technologien zu ermöglichen. Um speziell im 3D-Bereich validierte Interaktionspatterns durch eine formative und summative Evaluation intuitiver Benutzung bereitstellen zu können, muss als erstes spezifiziert werden, was unter einer intuitiven Benutzung überhaupt zu verstehen ist.

Dieses Kapitel gibt dazu zunächst einen Überblick über bestehende Definitionen intuitiver Benutzung aus dem Forschungsfeld der Mensch-Maschine-Interaktion und arbeitet daraufhin durch die Verknüpfung der Theorien von unterschiedlichen Forschergruppen mithilfe der Handlungsregulationstheorie als Metatheorie verschiedene charakterisierende Merkmale für dieses Konstrukt heraus. Auf Basis aktueller Forschung aus dem Bereich der Zwei-Prozess-Theorien werden anschließend die drei zentralen charakteristischen Merkmale identifiziert, mit denen zuverlässig das Ausmaß intuitiver Benutzung formativ und summativ evaluiert werden kann. Auf Grundlage des dadurch identifizierten subjektiven (d.h. metakognitives Gefühl von Flüssigkeit), des objektiven Merkmals (d.h. mentale Beanspruchung) und des pragmatischen Merkmals (d.h. Effektivität) wird abschließend eine eigene, für die Evaluation von intuitiver Benutzung passende, Arbeitsdefinition (d.h. Messdefinition) bereitgestellt.

2.1 Merkmale intuitiver Benutzung

Das Konzept der Intuition ist schon lange Forschungsgegenstand in Gebieten der Bildung und Lernpsychologie (z.B. Hogarth, 2001; Simonton, 1980), der Urteils- und Entscheidungsbildung (z.B. Evans & Stanovich, 2013; Kahneman & Frederick, 2002; Klein, 1993; Patterson, 2017), des kreativen Denkens (z.B. Bastick, 1982, 2003; K. S. Bowers, Farvolden, & Mermigis, 1995; Dorfman, Shames, & Kihlstrom, 1996; Mednick, 1962), der Kognitionswissenschaft (z.B. Damasio, 1994; Lieberman, 2000) und seit Neuestem im Bereich der Mensch-Maschine-Interaktion (z.B. Blackler, 2018; Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Hurtienne, 2011; O’Brien et al., 2010; Ullrich, 2014). Jeder Bereich betrachtet dabei Intuition aus einer anderen Perspektive. Intuition wird daher in den aufgezählten Bereichen nicht einheitlich verstanden, sondern stattdessen unter Berücksichtigung der Eigenheiten des jeweiligen Forschungsfeldes anhand bestimmter charakteristischer Merkmale eingegrenzt. Im weiteren Verlauf dieser Arbeit wird Intuition nun aus der Perspektive der Mensch-Maschine-Interaktion definiert und damit als *intuitive Benutzung* verstanden, die aus dem Wechselspiel von Aufgabe, Technologie und Nutzer resultiert (siehe Hurtienne, 2011). Wie in den anderen genannten Forschungsgebieten fällt im Bereich der Mensch-Maschine-Interaktion die Begriffsbestimmung aufgrund der inhärenten Komplexität des Begriffs schwer und explizite Definitionen für intuitive Benutzung, wie die der *IUUI-Forschergruppe* (d.h. Intuitive Use of User Interfaces) aus Deutschland

(z.B. Hurtienne, 2011; Mohs, Hurtienne, Kindsmüller et al., 2006; Naumann et al., 2007), der *QUT-Forschergruppe* (d.h. Queensland University of Technology) aus Australien (z.B. Blackler, 2006, 2018; Blackler & Popovic, 2015) und der *INTUI-Forschergruppe* (z.B. Diefenbach & Ullrich, 2015; Ullrich, 2014; Ullrich & Diefenbach, 2011) aus Deutschland bilden die grundlegenden Definitionen intuitiver Benutzung (Blackler, 2018). Diese Definitionen und die damit verbundenen charakteristischen Merkmale für intuitive Benutzung werden in den folgenden Teilabschnitten einzeln vorgestellt.

Da die unterschiedlichen Forschergruppen für die Ableitung dieser intuitiven Benutzung charakterisierenden Merkmale zum Teil unterschiedliche Theorien nutzen, wurde sich entschieden, diese Theorien mithilfe der Handlungsregulationstheorie von Hacker (1986) zu verknüpfen, um eine möglichst einheitliche Terminologie anhand dieser Metatheorie bereitstellen zu können. Hierzu werden im Folgenden die von den Forschergruppen genannten Merkmale intuitiver Benutzung nicht nur bezüglich der von den Forschergruppen verwendeten Literatur diskutiert, sondern diese Diskussion auch auf Basis von den Forschergruppen nicht explizit berücksichtigter Forschungsliteratur aus dem Bereich intuitiven Handelns (d.h. innerhalb und außerhalb der HCI) geführt. Auf diese Weise ist es möglich eine eindeutige Liste mit allen in der Literatur genannten Merkmalen (d.h. keine Duplikate, die zwar in verschiedenen Theorien unterschiedlich bezeichnet werden, aber das gleiche Merkmal beschreiben) und eine einheitliche Theorie als Grundlage dieser Liste von Merkmalen bereitzustellen, auf deren Basis im nächsten Abschnitt eine Arbeitsdefinition für die Evaluation intuitiver Benutzung (d.h. Messdefinition) abgeleitet werden kann.

2.1.1 IUUI-Forschergruppe

Die Forschergruppe *IUUI* bildete sich 2006 initiiert durch einige Teilnehmer des Graduiertenkollegs „prometei“ am Zentrum für Mensch-Maschine-Systeme an der Technischen Universität Berlin. Anhand eines Reviews der gängigsten Gestaltungskriterien zur Sicherstellung von Usability (siehe Scholz, 2006), der Analyse von Herstellerangaben bezüglich der intuitiven Benutzung ihrer Produkte, verschiedener Interviews und Workshops mit Nutzern sowie Usability-Experten konnte nach mehreren Verfeinerungen (siehe Hußlein et al., 2007; Israel et al., 2009; Mohs, Hurtienne, Kindsmüller et al., 2006; Naumann et al., 2008) folgende Definition intuitiver Benutzung der IUUI abgeleitet werden:

> Intuitive Benutzung ist das Ausmaß, mit dem ein Produkt durch die unbewusste Anwendung von Vorwissen genutzt wird. Sie führt zu effektiver und zufriedenstellender Interaktion, die ein Minimum an kognitiven Ressourcen beansprucht (Hurtienne, 2011).

Aus der IUUI-Definition können folgende vier Merkmale intuitiver Benutzung abgeleitet werden: die (1) *effektive Benutzung*, (2) die *mental effiziente Benutzung*, (3) die *zufriedenstellende Benutzung* und die (4) *unbewusste Anwendung von Vorwissen* während der Benutzung. Die ersten drei Merkmale stellen laut der IUUI-Forschergruppe dabei *Konsequenzen intuitiver Benutzung* dar und das letzte Merkmal bildet die *Grundvoraussetzung intuitiver Benutzung* (Hurtienne, 2011). Alle Merkmale können als Variablen (d.h. sie weisen eine bestimmte Variabilität auf), wie es durch den Begriff *Ausmaß* verdeutlicht wird, in der Stärke ihrer Ausprägung auf einem Kontinuum variieren (d.h. intuitive Benutzung

ist umso höher, je unbewusster, effektiver, mental effizienter und zufriedenstellender die Benutzung ist), weswegen intuitive Benutzung nicht als dichotomes Merkmal zu verstehen ist (Hurtienne, 2011). Die in der obigen Definition genannten Ausprägungen (z.B. effektiv) der Merkmale beschreiben somit eine vollkommen intuitive Benutzung (d.h. positives Extremum der jeweiligen Variable auf dem Kontinuum). Dabei treten laut obiger Definition alle Merkmale gleichzeitig im gleichen Ausmaß bei einer intuitiven Benutzung auf (d.h. alle Merkmale müssen in gleicher Ausprägung vorhanden sein, damit intuitive Benutzung vorliegt; intuitive Benutzung ohne Effektivität ist demnach beispielsweise nicht möglich) bzw. das Ausmaß intuitiver Benutzung spiegelt sich in allen Merkmalen gleichermaßen wider. Im Folgenden werden diese vier Merkmale unter Berücksichtigung der Handlungsregulationstheorie von Hacker (1986), diversen Arbeiten der IUUI-Forschergruppe und ergänzender Forschungsliteratur genauer spezifiziert.

2.1.1.1 Mental effiziente Informationsverarbeitung bei effektiver und zufriedenstellender Benutzung

Wie bereits in der Einleitung dieser Arbeit erwähnt, wird intuitive Benutzung oftmals als Subkonzept von Usability (d.h. Gebrauchstauglichkeit) verstanden, und dementsprechend auch von der IUUI-Forschergruppe so aufgefasst (Hurtienne, 2011; Naumann et al., 2007). Usability bzw. Gebrauchstauglichkeit wird innerhalb der DIN EN ISO 9241-11 definiert als „das Ausmaß, in dem ein Produkt durch bestimmte Benutzer in einem bestimmten Nutzungskontext genutzt werden kann, um bestimmte Ziele effektiv, effizient und zufriedenstellend zu erreichen“ (ISO, 1998, S. 4). Der Nutzungskontext ist demnach bestimmt durch „die Benutzer, Arbeitsaufgaben, Arbeitsmittel (z.B. Hardware, Software und Materialien) sowie die physische und soziale Umgebung, in der das Produkt genutzt wird“ (ISO, 1998, S. 4).

Mit der Anforderung intuitiver Benutzung an eine zufriedenstellende, mental effiziente und effektive Zielerreichung (d.h. Einbettung des Konzepts als Subkonzept innerhalb des Konstrukts der Usability) wird der Klassifizierung pseudo-intuitiver Handlungen als intuitive Handlungen vorgebeugt (siehe Ullrich, 2014). Laut Ullrich (2014) kann die Anforderung eines konkret formulierten Handlungsziels verhindern, dass ein Nutzer ein System als intuitiv beschreibt, weil er letztendlich effektiv (d.h. erfolgreich, aber nicht fehlerfrei) und mental effizient irgendwie per Trial-and-Error das System benutzen konnte und seine Interaktion deswegen als zufriedenstellend wahrnahm (z.B. zufälliges Finden eines Buches in einem Onlineshop anstelle zielgerichteten Suchens, was eventuell nicht so effektiv, mental effizient und zufriedenstellend sein kann). Es kann außerdem vorkommen, dass dieses Trial-and-Error-Verhalten nicht zielführend ist. Zusätzlich soll die Auffassung von intuitiver Benutzung als Subkonzept von Usability verhindern, dass intuitive Benutzung als dichotome Variable verstanden wird. Intuitive Benutzung kann wie Usability stattdessen als Variable auf einem Kontinuum bezüglich ihrer Merkmale in einem bestimmten Ausmaß variieren (Hurtienne, 2011). Aufgrund der Tatsache, dass es zwischen einem Nutzer und einem System immer zu einer Form von zielgerichteten Informationsaustausch im Zuge bestimmter Arbeitsaufgaben kommt (Naumann et al., 2007), kann intuitive Benutzung im Sinne der Handlungsregluationstheorie auch als Handlung verstanden werden (siehe Frese & Zapf, 1994; Hacker, 1986; Hacker & Sachse, 2013; Nerdinger, Blickle, Schaper,

& Schaper, 2014). Eine Handlung ist als zielgerichtetes Verhalten in einem bestimmten Nutzungskontext definiert. Bei einer Handlung steht für den Nutzer immer die eigene Nutzenmaximierung zu möglichst geringen mentalen Kosten im Vordergrund.

Tabelle 2.1. *Zusammenhänge zwischen den Konstrukten Usability und intuitiver Benutzung bezüglich der Leistungskriterien Effektivität, Effizienz und Zufriedenstellung anhand ausgewählter Indikatoren in Anlehnung an Hurtienne (2011).*

Usability-Leistungskriterien	Indikatoren	Relevanz dieser Leistungskriterien für intuitive Benutzung
Effektivität (Objektiv)	Fehlerrate Qualität der Zielerreichung Anteil erfolgreich erreichter Handlungsziele	Ja (beurteilen objektiv die Erreichung des Handlungsziels)
Effizienz (Objektiv)	Mentale Effizienz Anzahl an Hilfestellungen Benötigte Dokumentation Lernzeit	Ja (beziehen sich auf die kognitive Informationsverarbeitung)
	Physische Effizienz Zeit für motorische Handlungsausführung Finanzielle Kosten	Nein (beziehen sich nicht auf die kognitive Informationsverarbeitung)
Zufriedenstellung (Subjektiv)	Wahrgenommene Effektivität Wahrgenommene mentale Beanspruchung Präferenzen, Gefühle und Einstellungen	Ja (beurteilen subjektiv die Erreichung des Handlungsziels)

Dieser zielgerichtete Austausch, welcher auch als *Handlungsregulation* bezeichnet wird, findet solange in Form von Transformationen (d.h. Operationen) mit dem System statt, bis der Nutzer die Diskrepanz zwischen Soll-Zustand (d.h. vorgenommenes Handlungsziel) und Ist-Zustand (d.h. erhaltenes Handlungsergebnis von der Umwelt) überwunden und somit sein zuvor definiertes Handlungsziel erreicht hat (Hacker, 1986; Hacker & Sachse, 2013; Nerdinger et al., 2014; Zacher, 2017; Zempel, 2003). Mit dieser Rückmeldeschleife schließt sich der Handlungszyklus, weswegen eine Handlung oft auch als eine *zyklische Einheit* in Anlehnung an die TOTE-Einheit von G. A. Miller, Galanter und Pribram (1973) beschrieben wird. Hat der Nutzer sein Ziel erreicht, ist die Handlungsregulation beendet, andernfalls wird der Handlungszyklus solange iterativ durchlaufen, bis für den Nutzer eine *zufriedenstellende* Deckung zwischen Ist- und antizipierten Soll-Zustand erreicht ist (Frese & Zapf, 1994; Hacker, 1986; Hacker & Sachse, 2013; Zacher, 2017; Zempel, 2003). Usability und intuitive Benutzung teilen sich dementsprechend einige Gemeinsamkeiten aufgrund dieser Handlungsperspektive, wie die Tabelle 2.1 anhand der drei mit Usability verbundenen Leistungskriterien zeigt.

Ein erstes Merkmal intuitiver Benutzung stellt die damit verbundene Effektivität dar (Hurtienne, 2011; Mohs, Hurtienne, Kindsmüller et al., 2006; Naumann et al., 2007). Laut DIN EN ISO 9241-11 beschreibt Effektivität „die Genauigkeit und Vollständigkeit, mit der Benutzer ein bestimmtes Ziel erreichen“ (ISO, 1998, S. 4). Die Effektivität bei der Systemnutzung wird von beiden Konstrukten als objektives Leistungsmerkmal anerkannt und kann damit als Kriterium zur Beurteilung des Ausmaßes an intuitiver Benutzung fun-

gieren (siehe Tabelle 2.1). Falls Nutzer nämlich nicht in der Lage sind, ihre Handlungsziele mit angemessener Präzision und Vollständigkeit zu erreichen, wurde das System weder gebrauchstauglich noch intuitiv gestaltet (Hurtienne, 2011). Unter Berücksichtigung einer Handlungsperspektive entscheidet hierbei die Richtigkeit und Differenziertheit des bei der Handlung angewendeten Vorwissens über das Ausmaß der Effektivität, dessen Bewusstseinspflichtigkeit (d.h. bewusst vs. unbewusst) und allgemein die Güte der auf Basis des Vorwissens durchgeführten kognitiven Prozesse (Frese & Zapf, 1994; Hacker, 1986; Hacker & Sachse, 2013; Kauffeld, 2014).

Ergänzend kann hier noch Ulich (1994) genannt werden, der sagt, dass je realitätsangemessener (d.h. höherer Realitätsbezug durch hohe Richtigkeit und hohe Differenziertheit) das innere Modell während der Handlungsregulation ist, umso erfolgreicher kann die Arbeitstätigkeit ausgeführt werden. Die IUUI-Forschergruppe (siehe Hurtienne, 2011; Mohs, Hurtienne, Kindsmüller et al., 2006; Naumann et al., 2007) geht hier bereits einen nächsten logischen Schritt und erläutert, dass die unbewusste Anwendung von Vorwissen mit einer höheren Wahrscheinlichkeit zu einer effektiveren Interaktion führt. Im Folgenden soll jedoch im Sinne der Handlungsregulationstheorie die Richtigkeit und Differenziertheit des Vorwissens als Entscheidungsgrundlage betrachtet werden, obwohl die IUUI-Forschergruppe immer von unbewusster Anwendung von Vorwissen als Grundvoraussetzung intuitiver Benutzung spricht (siehe Hurtienne, 2011), da die unbewusste Anwendung des Vorwissen erst in Konsequenz aus der Richtigkeit und Differenziertheit des Vorwissens resultiert (siehe Frese & Zapf, 1994; Hacker, 1986; Hacker & Sachse, 2013; Nerdinger et al., 2014).

Hinsichtlich Effizienz lassen sich im Gegensatz zur Effektivität jedoch Unterschiede zwischen den beiden Konstrukten (d.h. intuitive Benutzung und Usability) feststellen (siehe Tabelle 2.1). Effizienz wird in der DIN EN ISO 9241-11 definiert als „der im Verhältnis zur Genauigkeit und Vollständigkeit eingesetzte Aufwand, mit dem Benutzer ein bestimmtes Ziel erreichen“ (ISO, 1998, S. 4). Sie steht aufgrund der Prämisse einer Nutzenmaximierung bei möglichst geringen ökonomischen Kosten, die jeder Handlung zugrunde liegt (Hacker, 1986; Hacker & Sachse, 2013; Nerdinger et al., 2014), zwar auch mit intuitiver Benutzung in Verbindung, jedoch stellen nur die mit der *kognitiven Informationsverarbeitung* in Verbindung stehenden objektiven Indikatoren (z.B. mentale Effizienz) auch Kriterien für die Beurteilung intuitiver Benutzung dar (siehe Tabelle 2.1). Diese Einschränkung liegt darin begründet, dass intuitive Benutzung sich aus Perspektive der IUUI-Forschergruppe als unbewusste Anwendung von Vorwissen versteht (Hurtienne, 2011; Mohs, Hurtienne, Kindsmüller et al., 2006; Naumann et al., 2007). Unter Berücksichtigung einer Handlungsperspektive bestimmt die Realitätsangemessenheit (d.h. Ausmaß an Richtigkeit und Differenziertheit) des bei der Handlung genutzten Vorwissens nicht nur über das Ausmaß an Effektivität (Hacker, 1986; Ulich, 1994), sondern auch über das Ausmaß an benötigter Bewusstseinspflichtigkeit bei der kognitiven Informationsverarbeitung während der Handlungsregulation (Frese & Zapf, 1994; Hacker, 1986; Hacker & Sachse, 2013). Kann die kognitive Informationsverarbeitung auf Basis realitätsangemessenen (d.h. richtigen und differenzierten) Vorwissens unbewusst erfolgen, so ist diese nicht nur effektiver, sondern auch gleichzeitig mental effizienter als die bewusste kognitive Informationsverarbeitung (siehe Hurtienne, 2011; Mohs, Hurtienne, Kindsmüller et al., 2006; Naumann et al., 2007). Die Anforderung intuitiver Benutzung nach mentaler Effizienz bei der kognitiven Informationsverarbeitung spiegelt damit auch den Anspruch von Handlungen nach Effizienz bei gleichzeitiger Nutzenmaximierung wider (siehe Hacker, 1986).

Im Forschungsfeld der Mensch-Maschine-Interaktion wird oftmals nicht nur die mentale Effizienz des Nutzers als subjektive Auswirkung auf den Nutzer thematisiert, sondern vielmehr die mentale Arbeitsbelastung, die während der kognitiven Informationsverarbeitung im Arbeitsgedächtnis auftritt, als Ganzes betrachtet (siehe Cain, 2007; F. Chen et al., 2016; Young, Brookhuis, Wickens, & Hancock, 2015). In Anlehnung an physische Arbeitsbelastung spiegelt sich mentale Arbeitsbelastung bei dieser Gesamtbetrachtung in zwei Komponenten wider: *Beanspruchung* und *Belastung* (Young et al., 2015). Unter mentaler Belastung versteht man hier die kognitiven Anforderungen an das Arbeitsgedächtnis (d.h. *Anforderungsperspektive*) durch die Aufgabe und Umgebung, welche extern auf den Nutzer einwirken, aber von diesem nicht beeinflusst werden können. Es existieren eine Reihe von Belastungsfaktoren, die aufgrund der Umwelt (z.B. Lärm, Beleuchtung) und der konkreten Arbeitsaufgabe (z.B. Komplexität der Informationen, verwendete Bedienelemente und Darstellungsart) auf den Nutzer wirken und seine kognitive Informationsverarbeitung beeinflussen (Schlick, Winkelholz, Motz, Duckwitz, & Grandt, 2010).

Die mentale Beanspruchung, was der mentalen Effizienz des Nutzers entspricht, die im Rahmen des Konstrukts der Usability (damit auch intuitiver Benutzung) objektiv und subjektiv als Variable bewertet wird (siehe Tabelle 2.1), beschreibt die individuelle Auswirkung dieser kognitiven Faktoren auf das Arbeitsgedächtnis des Nutzers (d.h. *Wirkungsperspektive*). Dabei hängt die Beanspruchung, die jeder Nutzer individuell erlebt, neben den objektiven Faktoren noch von verschiedenen weiteren subjektiven Faktoren ab (z.B. individuellem Vorwissen, Frustration, Motivation, Zufriedenheit). Diese Faktoren nehmen dabei gleichzeitig Einfluss auf die kognitive Informationsverarbeitung im Arbeitsgedächtnis und die dafür zur Verfügung stehenden mentalen Ressourcen (Galy, Cariou, & Mélan, 2012; Galy, Paxion, & Berthelon, 2018; Young et al., 2015). Auf Grundlage dieser im Forschungsfeld der Mensch-Maschine-Interaktion vorherrschenden Auffassung, kann mentale Effizienz bzw. mentale Beanspruchung in Anlehnung an Young et al. (2015) als das Ausmaß an benötigten kognitiven Ressourcen des Arbeitsgedächtnisses (d.h. Kurzzeitgedächtnisses) definiert werden, die unter Berücksichtigung des Nutzungskontextes zur effektiven Erfüllung des Handlungsziels nötig sind. Der Begriff Arbeitsgedächtnis wird im nächsten Teilabschnitt dieses Abschnitts detailliert erläutert. Mentale Beanspruchung und mentale Effizienz sind im weiteren Verlauf dieser Arbeit als Synonyme zu sehen. Wenn von mentaler Arbeitsbelastung gesprochen wird, ist fortan immer das Gesamtkonzept gemeint, was sich aus den beiden vorgestellten Komponenten zusammensetzt (siehe Young et al., 2015).

Die physische Effizienz während der Handlungsregulation oder die zeitliche Effizienz bei der letztendlichen motorischen Ausführung der Handlung korrelieren, im Gegensatz zu mentaler Effizienz, jedoch nicht notwendigerweise mit intuitiver Benutzung (siehe Tabelle 2.1), da eine körperlich anstrengende oder langsame Bewegungssequenz im Rahmen einer Handlung keine intuitive Handlung impliziert (z.B. ein Installationswizard, welcher einem Nutzer viele Mausklicks abverlangt, ist dadurch nicht automatisch weniger intuitiv, da dies nicht die kognitive Informationsverarbeitung beeinträchtigt) (Hurtienne, 2011; Naumann et al., 2007). An dieser Stelle soll jedoch schon einmal vorweggenommen werden, dass die genannte zeitliche Effizienz mit einer Einschränkung trotzdem als Kriterium zur Beurteilung intuitiver Benutzung herangezogen werden kann. Nur die gesamte Durchführungszeit von Handlungen (d.h. zeitliche Effizienz bei der Handlungsdurchführung) sollte theoretisch nur gering mit dem Ausmaß an intuitiver Benutzung korrelieren, da diese auch die Ausführungszeiten der motorischen Handlungsausführung beinhaltet und nicht nur die Zeit

für die kognitive Informationsverarbeitung selbst, die bei intuitiven Handlungen relativ kurz ist. Aufgrund der Tatsache, dass intuitive Benutzung nur die kognitive Informationsverarbeitung und damit die mentale Beanspruchung widerspiegeln kann, korreliert sie auch nur hoch mit der damit zusammenhängenden zeitlichen Effizienz (d.h. Dauer für die kognitive Informationsverarbeitung). Eine mental effiziente Informationsverarbeitung ist demnach auch zeitlich effizient, da die Informationen mithilfe eines hohen realitätsangemessenen (d.h. richtigen und differenzierten) Vorwissens auch schnell verarbeitet werden können (siehe Blackler, 2006, 2018; Evans & Stanovich, 2013; Greenfield, 2000; R. Reber & Schwarz, 2001; R. Reber, Wurtz, & Zimmermann, 2004; Topolinski, 2011). Auf diesen Aspekt wird im kommenden Teilabschnitt 2.1.2 vertieft eingegangen, da die IUUI-Forschergruppe diesen Aspekt in ihrer Definition nicht als weiteren Zugang zu mentaler Effizienz explizit berücksichtigt. Wahrscheinlich, weil die Zeit für die Handlungsausführung und die Zeit für die kognitive Informationsverarbeitung operativ schwer zu trennen sind.

Letztlich bildet die subjektive Zufriedenstellung oder Zufriedenheit aus Sicht der IUUI-Forschergruppe eine weitere Konsequenz intuitiver Benutzung (siehe Hurtienne, 2011; Naumann & Hurtienne, 2010; Naumann et al., 2007), welches neben Effektivität und mentaler Effizienz als weiteres Leitkriterium für Usability im Rahmen der DIN EN ISO 9241-11 berücksichtigt ist (ISO, 1998). Zufriedenstellung wird in der Norm definiert als die „Freiheit von Beeinträchtigungen und positive Einstellung gegenüber der Nutzung des Produkts“ (ISO, 1998, S. 4). Diese subjektive Einschätzung des Ergebnisses der Handlungsregulation (z.B. erfasst durch das Verhältnis von positiven zu negativen Nutzerkommentaren) ist zwar für beide Konstrukte gleichermaßen wichtig, wobei sich für eine Beurteilung der Handlungsregulation bezüglich intuitiver Benutzung vorwiegend Indikatoren eignen, die mit der kognitiven Informationsverarbeitung zusammenhängen. Diverse Studien zeigen, dass die bewusste kognitive Informationsverarbeitung nicht nur weniger effektiv und weniger mental effizient ist, sondern auch zu weniger Zufriedenheit führt als eine unbewusste kognitive Informationsverarbeitung (siehe Betsch, Funke, & Plessner, 2011; Dijksterhuis & Van Olden, 2006; Ullrich, 2014).

Die erhöhte Zufriedenheit aufgrund der unbewussten Anwendung von Vorwissen drückt sich laut der IUUI-Forschergruppe in den folgenden fünf subjektiven Indikatoren aus (Hurtienne, 2011; Naumann & Hurtienne, 2010):

Wahrgenommene geringe mentale Beanspruchung leitet sich direkt aus dem Merkmal einer mental effizienten Benutzung ab (Hurtienne, 2011; Naumann & Hurtienne, 2010).

Wahrgenommene hohe Zielerreichung leitet sich direkt aus dem Merkmal einer effektiven Benutzung ab (Hurtienne, 2011; Naumann & Hurtienne, 2010).

Wahrgenommene geringe Fehlerrate leitet sich direkt aus dem Merkmal einer effektiven Benutzung ab (Hurtienne, 2011; Naumann & Hurtienne, 2010).

Wahrgenommene hohe Vertrautheit leitet sich indirekt aus der unbewussten Anwendung von Vorwissen ab, da diese zu einem Gefühl von Vertrautheit bei den Nutzern führt (Hurtienne, 2011; Naumann & Hurtienne, 2010).

Wahrgenommener geringer Lernaufwand leitet sich indirekt aus der unbewussten Anwendung von Vorwissen ab, da richtiges und differenziertes Vorwissen unbewusst für

die Interaktion genutzt werden kann. Der Lernaufwand des Nutzers bleibt infolgedessen vergleichsweise gering (Hurtienne, 2011; Naumann & Hurtienne, 2010). Darüber hinaus merken weitere, nicht mit der IUUI-Forschergruppe assoziierte Wissenschaftler wie Reddy (2012) und O'Brien, Rogers und Fisk (2008) an, dass auch bei noch nicht vollständig richtigem und differenziertem Vorwissen das System „nachsichtig" gestaltet sein soll (z.B. durch sofortiges Feedback, Undo-Funktion), sodass richtiges und differenziertes Vorwissen implizit mit geringem Aufwand durch den Nutzer aufgebaut werden kann.

Da die Realitätsangemessenheit (d.h. Richtigkeit und Differenziertheit) des Vorwissens für den jeweiligen Nutzungskontext offenbar als Grundvoraussetzung intuitiver Benutzung darüber entscheidet, ob es unbewusst oder bewusst kognitiv verarbeitet werden kann, was sich dann wiederum in den objektiven und subjektiven Konsequenzen intuitiver Benutzung (d.h. Effektivität, mentaler Effizienz, Zufriedenstellung) entsprechend gleichermaßen positiv ausgeprägt auf einem Kontinuum widerspiegelt (siehe Hurtienne, 2011; Naumann et al., 2007), soll diese Grundvoraussetzung intuitiver Benutzung im nächsten Abschnitt genauer charakterisiert werden.

2.1.1.2 Unbewusste Anwendung von Vorwissen

Der Prozess der Handlungsregulation stellt für den Nutzer einen komplexen Prozess dar, bei dem der Nutzer in einem kognitionspsychologischen Sinn als Informationsverarbeitungseinheit aufgefasst wird. Als Informationsverarbeitungseinheit sammelt dieser dann bezüglich seines Handlungsziels aktiv Informationen aus seiner Umgebung, speichert diese, fügt diese mit seinem Vorwissen zu einem mentalen Modell zusammen und wendet dieses Modell dann zur Lösung bestimmter Probleme (d.h. Barrieren der Handlungsregulation) bei der Systemnutzung an (siehe Diefenbach & Hassenzahl, 2017; Nerdinger et al., 2014; Saifoulline & Hemberger, 2011). Dabei unterscheiden alle gängigen Informationsverarbeitungsmodelle (siehe Baddeley, 2012; A. Newell, Simon et al., 1972; Vollrath, 2015; Wickens, Hollands, Banbury, & Parasuraman, 2015) zwischen drei mnestischen Systemen (d.h. sensorisches Gedächtnis, Kurzzeitgedächtnis, Langzeitgedächtnis), sowie Antwortgenerator und zentralem Prozessor (Nerdinger et al., 2014). Die kognitive Verarbeitung von Informationen erfolgt dabei vereinfacht zunächst mithilfe des sensorischen Gedächtnisses (d.h. Rezeptorsystem, sensorisches Register, Ultrakurzzeitgedächtnis), das visuelle, auditive oder haptisch physische Informationen aufnimmt und dann an einen zentralen Prozessor (d.h. zentrale Exekutive) weiterleitet. Dieser Prozessor führt nach einem bestimmten Ablaufschema eine Reihe von elementaren Operationen zur Kodierung, Verarbeitung und Speicherung durch. Dazu nutzt er die beiden anderen mnestischen Gedächtnissysteme: Langzeit- und Kurzzeitgedächtnis (d.h. Arbeitsgedächtnis) (Nerdinger et al., 2014; A. Newell, Simon et al., 1972; Vollrath, 2015; Wickens et al., 2015).

Das Langzeitgedächtnis, welches das gesamte Vorwissen des Nutzers enthält, übernimmt vor allem die langfristige Speicherung von Informationen in Form von mentalen Wissensrepräsentationen. Es besitzt virtuell unbegrenzte Kapazität (Nerdinger et al., 2014), jedoch ist der Zugriff (d.h. sich erinnern) auf diese Wissensrepräsentationen mit 0.5 Sekunden relativ langsam (Atkinson & Shiffrin, 1968). Der Begriff Vorwissen wurde im Bereich der Lernpsychologie ursprünglich von Jonassen und Grabowski (1993) geprägt. Er wird dort

als das gesamte tatsächliche Wissen einer Person im Langzeitgedächtnis verstanden, was für die kognitive Informationsverarbeitung benötigt wird. Dieses Vorwissen ist dynamisch (d.h. optimierbar durch Intensität und Häufigkeit der Anwendung, siehe Bargh & Chartrand, 1999), steht vor der Beginn einer Handlung zur Verfügung, ist strukturiert (d.h. inhaltlich und strukturell durch Handlungsziele determiniert), liegt in unterschiedlichen mentalen Repräsentationsformen vor, ist zum Teil explizit, zum Teil implizit und umfasst auch metakognitive Komponenten (Dochy & Alexander, 1995; Dochy, Segers, & Buehl, 1999; Krause & Stark, 2006). Diese Dimensionen des Vorwissens mit deren Hilfe dieses kategorisiert werden kann, sollen nun nachfolgend kurz vorgestellt werden. Auf den metakognitiven Aspekt des Vorwissens wird jedoch erst in Teilabschnitt 2.1.3 kurz und in Abschnitt 2.2 ausführlich Bezug genommen, da solche Aspekte nicht von der IUUI-Forschergruppe thematisiert werden und hier vorwiegend die theoretische Grundlage der IUUI-Forschergruppe erläutert werden soll.

Bezüglich der inhaltlichen Struktur lässt sich zwischen deklarativem (d.h. Wissen über Fakten, Bedeutungen von Symbolen, Konzepte und Prinzipien eines bestimmten Nutzungskontextes bzw. einer Domäne), prozeduralem Wissen (d.h. Wissen über Handlungen und Fertigkeiten) und konditionalem Wissen (d.h. Wissen darüber, in welcher Situation bestimmtes Wissen zu Anwendung kommt) unterscheiden (Krause & Stark, 2006). Darüber hinaus lässt sich zwischen domänenspezifischem Wissen (d.h. Expertenwissen) und domänenübergreifendem Wissen (d.h. universellem Wissen) unterscheiden (Krause & Stark, 2006). Außerdem kann Vorwissen explizit oder implizit vorliegen und sich damit bezüglich der Bewusstseinsdimension unterscheiden. Explizites Wissen ist dabei formal verbalisierbar und kann bewusst aktiviert werden. Implizites Wissen wird unbewusst aktiviert und kann dementsprechend nicht oder nur formal unvollständig in sprachlicher Form wiedergeben werden. Prozedurales Wissen ist überwiegend implizit und deklaratives explizit (Krause & Stark, 2006).

Vorwissen ist im Langzeitgedächtnis unterschiedlich repräsentiert und bildet die Grundvoraussetzung für die kognitive Informationsverarbeitung im Rahmen einer jeden Handlungsregulation und damit auch für eine damit verbundene intuitive Benutzung. Es existieren eine Reihe von kognitionspsychologischen Modellen, die unterschiedliche Annahmen über die Vorwissensaktivierung bzw. den Abruf des Vorwissens während der Handlungsregulation treffen. Im Rahmen dieser Arbeit soll nur auf Schemas als mentales Repräsentationsmodell eingegangen werden, da diese Art als generelle Repräsentation von Vorwissen (siehe D. A. Norman & Shallice, 1986; Rumelhart, 2017; Sowa, 2000) von der IUUI-Forschergruppe bereits eingeführt wurden (siehe Hurtienne, 2011) und Schemas auch unabhängig davon als Repräsentationsart im Vergleich zu anderen in der Literatur genannten Repräsentationsmodellen am weitesten verbreitet sind (siehe Brewer & Nakamura, 1984; Rumelhart, 2017; Sowa, 2000). Für weitere mentale Repräsentationsmodelle wird auf Krause und Stark (2006) verwiesen.

Schemas sind abstrakte mentale Wissensrepräsentationen und dienen als kategorisierende, sowie generalisierende Zusammenfassungen von Objekten und Handlungen, abgesehen von ihren individuellen Charakteristika unter Bewahrung ihrer gemeinsamen Kernaussage (Brewer & Nakamura, 1984; Hurtienne, 2011; Kopp & Mandl, 2005; Rumelhart, 2017). Laut Anderson und Funke (2001) können Schemas deswegen für Schlussfolgerungen über

spezifische Objekte oder Handlungen genutzt werden. Sie können ineinander geschachtelt sein und weisen eine hohe Flexibilität auf, da sie Vorwissen auf unterschiedlichen Abstraktionsebenen repräsentieren können (Hurtienne & Blessing, 2007; Kopp & Mandl, 2005; Rumelhart, 2017). Auf einer niedrigen Abstraktionsebene befinden sich motorische Handlungsschemas (d.h. motor response schemas; sensomotorische Schemas), die die Zusammenhänge zwischen den Ausgangsbedingungen einer Bewegung und dem gewünschten Ergebnis in einer abstrakter Form (z.B. Schema für Laufen, Schema für Greifen) ausdrücken (R. Cooper & Shallice, 2000; Hurtienne, 2011). Motorische Handlungsschemas werden von Kindern als erstes durch sensorische Verarbeitung von visuellen und taktilkinästhetischen Informationen generiert (Piaget, 1976).

Im Vergleich zu motorischen Handlungsschemas sind Image Schemas auf einer höheren Abstraktionsebene zu finden (Hurtienne, 2011). Sie verstehen sich als eine sensomotorische abstrakte Form unbewusster Wissensrepräsentation, die sich durch wiederkehrende sensomotorische Erfahrungen mit der Umwelt im Alltag bildet (Johnson, 2013). Mithilfe des kognitiven Vorgangs der perzeptuellen Bedeutungsanalyse werden hier wahrgenommene Informationen in konzeptuelles, abstraktes Wissen (d.h. Bedeutung) überführt, dessen mentale Repräsentation selbst nicht wahrnehmungsgebunden ist (Mandler, 1992). Johnson (2013) beschrieb ursprünglich 29 solcher Schemas. Image Schemas sind abstrakter als motorische Handlungsschemas und umfassen mehr als die reine sensomotorische Wahrnehmung, da sie auch Objekt- und Ereignischarakteristika mitberücksichtigen, die sensomotorisch erfasst wurden (Hurtienne, 2011). Der Unterschied kann anhand des folgenden Beispiels verdeutlicht werden. Ein Legostein kann im Sinne eines motorischen Handlungsschema als etwas Greifbares wahrgenommen werden (d.h. Greif-Schema). Im Sinne eines Image Schemas kann darüber hinaus wahrgenommen werden, ob sich der Legostein in der Nähe befindet (d.h. Image Schema NEAR-FAR) oder bereits mit anderen Legosteinen zu etwas Größerem verbaut wurde (d.h. Image Schema PART-WHOLE). Kinder stellen bereits in ihrem ersten Lebensjahr fest, dass sich Objekte räumlich in der Nähe oder getrennt voneinander befinden können und so eine gewisse Zugehörigkeit symbolisiert wird (Mandler, 1992, 2005, 2014; G. A. Miller & Johnson-Laird, 1976). Für eine detaillierte Auseinandersetzung mit Schemas wird auf Cienki und Müller (2008), Cienki (2015), Mandler (2014) und Hurtienne (2017) verwiesen.

Vorwissen besitzt neben den bereits vorgestellten Dimensionen (z.B. Inhalt, Bewusstheit, Repräsentation) außerdem noch Dynamik, da es sich aus unterschiedlichen Wissensquellen zusammensetzt und durch intensive, sowie häufige Anwendung mit höherer Wahrscheinlichkeit unbewusst zugänglich gemacht werden kann. Einiges Wissen (z.B. Reflexe) ist dabei bereits von vornherein unbewusst zugänglich (Bargh & Chartrand, 1999; Hurtienne, 2011). Laut der IUUI-Forschergruppe können diese Wissensrepräsentationen bei der Systemnutzung aus verschiedenen Wissensquellen stammen, die mithilfe des Kontinuums der Wissensquellen (siehe Abbildung 2.1) untersucht werden können (Hurtienne, 2011).

Auf der untersten Ebene, welche als universelle Wissensebene angesehen werden kann, befindet sich angeborenes Wissen (z.B. Instinkte und Reflexe) (Hurtienne, 2011). Die nächste Ebene bildet das sensomotorische Wissen. Es besteht aus Allgemeinwissen, welches früh in der Kindheit erworben und seitdem kontinuierlich durch Interaktionen mit der eigenen Umwelt durch lernendes Handeln verinnerlicht wurde (Hurtienne, 2011). Beispielsweise lernen Kinder recht früh verschiedene Gesichter zu unterscheiden und wie Schwerkraft

funktioniert. Auf der nächsten Ebene befindet sich kulturabhängiges Wissen (Hurtienne, 2011). Wissen, welches im Umgang mit westlichen Kulturen erworben wurde, muss nicht notwendigerweise auch bei östlichen Kulturen erfolgreich anwendbar sein (z.B. erwünschte Farbe bei Beerdigungen oder Hochzeiten). Die spezifischste Ebene des Wissens ist Expertenwissen, welches man sich im Rahmen seiner Profession (z.B. Wissenschaftler, Schneider) oder aufgrund seiner Hobbys (z.B. Motorrad fahren, Ballett) angeeignet hat (Hurtienne, 2011). Auf der sensomotorischen, kulturellen und fachlichen Ebene kann das Wissen durch den Umgang mit Werkzeugen und Technologien (z.B. Verwendung eines CAD-Programms) gewonnen werden (Hurtienne, 2011). Solches Wissen spielt eine übergeordnete Rolle bei der Benutzung von Mensch-Maschine-Systemen und kann als übergreifende Wissensebene bei deren Benutzung verstanden werden (Hurtienne, 2011; Hurtienne & Blessing, 2007; Naumann et al., 2007).

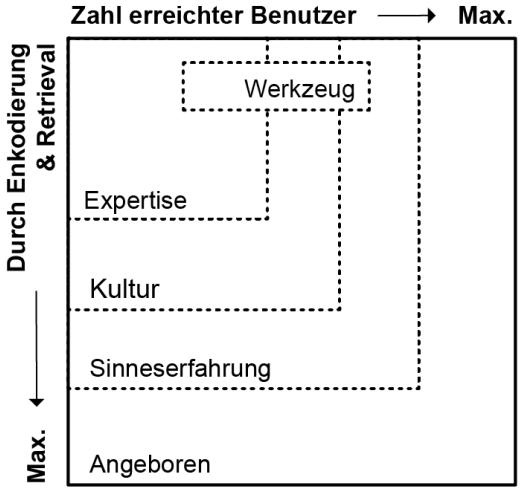


Abbildung 2.1. Kontinuum der Wissensquellen in Anlehnung an Hurtienne (2011).

Das Kontinuum der Wissensquellen besitzt außerdem eine inhärente Dimension (Blackler & Hurtienne, 2007; Hurtienne, 2011), um seine Dynamik und damit eine unterschiedliche Wahrscheinlichkeit bezüglich dessen Richtigkeit und Differenziertheit für bestimmte Handlungssituationen vorhersagen zu können. Je weiter unten sich die Wissensquelle im Kontinuum befindet, umso wahrscheinlicher kann dieses Wissen von vielen verschiedenen Nutzern in verschiedenen Nutzungskontexten (siehe horizontale Dimension in Abbildung 2.1) unbewusst angewendet werden, da Nutzer ihr Wissen durch viele ähnliche Erfahrungen in diversen Nutzungskontexten bereits gut differenzieren und vervollständigen konnten. Im Gegensatz dazu kann beispielsweise später im Leben erworbenes Expertenwissen schlechter auf andere Kontexte übertragen werden, da dieses Wissen nicht übergreifend für eine Vielzahl von Handlungen relevant ist. Jedoch kann Vorwissen unabhängig davon, wo es sich auf dem Kontinuum befindet durch eine intensive und häufige Anwendung (d.h. Übung) in seiner Richtigkeit und Differenziertheit für diverse Nutzungskontexte optimiert (d.h. Vorwissen ist mit hoher Wahrscheinlichkeit richtig und differenziert genug, um in vielen Kontexten effektiv angewendet zu werden) und dadurch mit höherer Wahrscheinlichkeit unbewusst angewendet werden (siehe vertikale Dimension in Abbildung 2.1). Der

Aufwand für die Optimierung des Vorwissens ist umso höher, je spezifischer das Wissen bzw. je weiter oben es sich im Kontinuum der Wissensquellen befindet (Hurtienne, 2011).

In der Terminologie von Informationsverarbeitungsmodellen und der damit verbundenen Auffassung, dass es sich beim Nutzer aus einer kognitionspsychologischen Perspektive um eine Informationsverarbeitungseinheit handelt (Nerdinger et al., 2014; A. Newell, Simon et al., 1972; Vollrath, 2015; Wickens et al., 2015), stellt die Anwendung des Vorwissens bzw. die Vorwissensaktivierung den Abruf gespeicherter Informationen aus dem Langzeitgedächtnis und die Bereitstellung dieser Informationen im Kurzzeit- bzw. Arbeitsgedächtnis dar (siehe Krause & Stark, 2006). Das Arbeitsgedächtnis ist im Wesentlichen für die zeitlich relativ begrenzte Bereitstellung von Informationen zur Verarbeitung durch den zentralen Prozessor zuständig und gilt als das einzige mnestische System, dessen Inhalte der bewussten Verarbeitung (d.h. kognitiven Denkens) zugänglich sind (Nerdinger et al., 2014). Eines der bis heute einflussreichsten Modelle des Arbeitsgedächtnisses, welches zunächst aus den drei Komponenten *phonologische Schleife*, *räumlich-visueller Notizblock* und *der zentralen Exekutive* als zentraler Prozessor bestand, lieferten Baddeley und Hitch (1974). Baddeley (2000) erweiterte später das Modell um den *episodischen Puffer*, der es der zentralexekutiven Steuerung erlaubt Informationen verschiedener Modalitäten aus dem Langzeitgedächtnis, dem visuell-räumlichen Notizblock und der phonologischen Schleife zu komplexen einheitlichen episodischen, multimodalen mentalen Repräsentationen (z.B. Schemas) zu verknüpfen.

Das erste Subsystem, die phonologische Schleife, erlaubt es gesprochene und geschriebene Informationen im Gedächtnis in Lautform (d.h. Phoneme) zu bewahren und zu verwalten (Baddeley, 2012; J. D. Lee, Wickens, Liu, & Boyle, 2017). Die Schleife besteht selbst aus zwei Komponenten, einem passiven phonetischen Speicher und einem aktiven subvokalen artikulatorischen Kontrollprozess. Mithilfe des Speichers können auditorisch-verbale Informationen durch eine phonologische Repräsentation für etwa 1.8 Sekunden bewahrt werden, während es der subvokale Rehearsalprozess gestattet, dass auf die gesprochenen Informationen auch über diese Zeitspanne hinaus zugegriffen werden kann (Baddeley, 2012; Hasselhorn & Grube, 2003). Zum Beispiel findet die phonologische Schleife Anwendung, wenn sich Menschen für kurze Zeit eine Telefonnummer merken müssen. Wir sprechen uns innerlich solange die Nummer vor, bis wir sie entweder in die Tastatur des Telefons eingeben oder sie uns gemerkt haben (J. D. Lee et al., 2017). Der Kontrollprozess erlaubt ferner eine phonetische Umkodierung visuell dargebotener Informationen (Hasselhorn & Grube, 2003).

Das zweite Subsystem, der visuell-räumliche Notizblock, ist auf die Speicherung und Verwaltung von visuellen und räumlichen Informationen spezialisiert (J. D. Lee et al., 2017). Ferner kann auch verbales Material damit verarbeitet werden, sofern dieses vom Subsystem in eine analoge räumliche Form (z.B. Bildsprache, siehe Logie, 2014) umkodiert wird (Hasselhorn & Grube, 2003). In Analogie zur phonologischen Schleife setzt sich dieses Subsystem ebenfalls aus zwei Komponenten zusammen, einem passiven visuellen Speicher und einem aktiven räumlichen Rehearsalprozess. Der visuelle Zwischenspeicher speichert visuell wahrgenommenes Material in Form und Farbe und hält diese Information für einige Zeit aufrecht. Der Rehearsalprozess verarbeitet Positions- und Bewegungsinformationen, welche nur über diesen Prozess in den visuellen Zwischenspeicher gelangen können (Baddeley, 1999, 2012; Logie, 2011, 2014; Quinn & McConnell, 1996). Des Weiteren transferiert der

Rehearsalprozess Informationen vom Zwischenspeicher zur zentralen Exekutive. Es werden im Allgemeinen durch den Rehearsalprozess mehr mentale Ressourcen verbraucht als dies beim Rehearsalprozess der phonologischen Schleife der Fall ist (Logie, 2011). Es soll an dieser Stelle jedoch angemerkt werden, dass die Forschung bezüglich dieses zweiten Subsystems nicht in dem Maße fortgeschritten ist, wie es bei der phonologischen Schleife der Fall ist. Wie genau das Rehearsal beim visuell-räumlichen Notizblock genau funktioniert ist demnach noch nicht vollständig erforscht (Baddeley, 2012).

Das dritte Subsystem, der episodische Puffer, erlaubt es Informationen aus dem visuell-räumlichen Notizblock, der phonologischen Schleife und dem Langzeitgedächtnis zu integrieren und als kohärente Episoden (d.h. chunks of information, siehe Gooding, Isaac, & Mayes, 2005) zu speichern. Das System kommuniziert dabei mit der phonologischen Schleife, dem Langzeitgedächtnis und dem visuell-räumlichen Notizblock, um die dort enthaltenen Informationen semantisch zu interpretieren (Baddeley, 2000, 2001, 2012). Die Kombination der Inhalte erfolgt dabei anhand der aus dem Langzeitgedächtnis abgerufenen Schemas, die beschreiben, inwiefern Informationen zu sinnvollen Einheiten kombiniert werden können (Gooding et al., 2005). Das System ist dabei temporär und kapazitär beschränkt (Baddeley, 2000, 2001, 2012) und kann nur 7 ± 2 gebildete Informationseinheiten gleichzeitig enthalten (G. A. Miller, 1956). Das System vermittelt der exekutiven Kontrolle den bewussten Zugang zu diesen kohärenten Informationen (Baddeley, 2012).

Bei der zentralen Exekutive handelt es sich um den zentral überwachenden Prozessor (d.h. das eigentliche Arbeitsgedächtnis), welchem die Verantwortung für Regulations- und Kontrollaktivitäten zugeschrieben wird. Auf diese Weise ermöglicht sie dem Menschen eine zielgerichtete Handlungsregulation (Baddeley, 2012). Dabei koordiniert die zentrale Exekutive im Wesentlichen die drei vorgestellten Subsysteme des Arbeitsgedächtnisses (d.h. phonologische Schleife und visuell-räumlicher Notizblock als sensorische Eingänge und den episodischen Puffer als die Verbindung zwischen dem Arbeitsgedächtnis und dem Langzeitgedächtnis), sowie den Abruf und die Manipulation von mentalen Repräsentationen (d.h. Schemas) aus dem Langzeitgedächtnis (Baddeley, 2012). Der zentralen Exekutive werden im Rahmen der Handlungsregulation eine Reihe weiterer Funktionen zugeordnet, wie beispielsweise das Setzen und Priorisieren von Handlungszielen, die strategische Handlungsplanung, deren Umsetzung und die Kontrolle der motorischen Umsetzung (Jurado & Rosselli, 2007). Diese Aufzählung ist nur exemplarisch zu verstehen. Für einen vollständigeren Überblick wird auf Jurado und Rosselli (2007) verwiesen.

Im Rahmen dieser Arbeit soll sich auf die drei basalen Grundfunktionen der zentralen Exekutive beschränkt werden, da sie allen anderen höheren exekutiven Funktionen (z.B. Setzen von Handlungszielen) zugrunde liegen (Baddeley, 2007; Friedman & Miyake, 2004; Logie, 2011; A. Miyake et al., 2000; Nee et al., 2012; Repovš & Baddeley, 2006). Bei den drei Grundfunktionen handelt es sich um die *Inhibition der unbewussten Reaktion* (d.h. inhibitorische Kontrolle), die *Aktualisierung von mentalen Repräsentationen im Arbeitsgedächtnis* (d.h. Updating) und die *Durchführung des Wechsels zwischen Aufgaben, sowie des Wechsels zwischen bewusster und unbewusster kognitiver Informationsverarbeitung* (d.h. Shifting). Aktuelle Forschung geht zwar davon aus, dass man zwar theoretisch zwischen diesen drei exekutiven Grundfunktionen unterscheiden kann, diese aber immer miteinander korrelieren (A. Miyake et al., 2000). Dementsprechend kann die exekutive

Kontrolle auch als einheitliches System betrachtet werden, an dem sich die mentale Beanspruchung des Nutzers ablesen lässt, da sich seine Funktionen gegenseitig beeinflussen und bedingen (Jurado & Rosselli, 2007).

Den drei exekutiven Grundfunktionen werden in Forschungsliteratur die folgenden Aufgaben zugeordnet:

Inhibition beschreibt die Fähigkeit impulsive unbewusste Handlungstendenzen zu kontrollieren oder zu hemmen, um durch logisches Denken und damit verbundene bewusste Kontrolle Entscheidungen zu treffen (A. Miyake et al., 2000). Diese Fähigkeit zur Inhibition erlaubt uns weniger geläufige, aber zielführende Handlungen in unbekannten Situationen zu zeigen. Dadurch impliziert der Begriff der inhibitorischen Kontrolle, das Filtern von Information und das Festlegen von Prioritäten bei der Bewusstseinssteuerung.

Updating beschreibt die Fähigkeit zur Überwachung und Kodierung neu eingehender Informationen. Es wird dabei immer geprüft, inwiefern die Informationen für die aktuelle Handlung relevant sind und abhängig vom Ergebnis entschieden, ob relevante Informationen im Arbeitsgedächtnis durch Rehearsal-Prozesse aufrechterhalten bzw. irrelevante Informationen durch neue ersetzt werden. Die Aktualisierung der Wissensrepräsentationen im Arbeitsgedächtnis geht über die rein passive Speicherung und Aufrechterhaltung im Kurzzeitgedächtnis hinaus, da die beteiligten Prozesse auch die aktive Manipulation von Information ermöglichen (A. Miyake et al., 2000).

Shifting erlaubt es dem Menschen zwischen parallel durchgeführten Handlungen, verschiedenen Operationen und der Art der kognitiven Informationsverarbeitung (d.h. unbewusst oder bewusst) dynamisch hin und her wechseln zu können (A. Miyake et al., 2000).

Aus der komplexen Verbindung der drei mnestischen Systeme resultiert das handlungsrelevante Vorwissen als *mentales Modell des Nutzungskontextes*, welches im Rahmen der Handlungsregulationstheorie auch als operatives Abbildsystem bezeichnet wird (Frese & Zapf, 1994; Hacker, 1986; Hacker & Sachse, 2013; Nerdinger et al., 2014). Dieses mentale Modell erlaubt dem Nutzer auf Basis seines Vorwissens eine mehr oder weniger genaue Vorstellung von der Systemnutzung (Frese & Zapf, 1994) und ermöglicht ihm zielgerichtetes Handeln auf Basis seiner exekutiven Funktionen. Es fungiert damit als Grundvoraussetzung für das Handeln im Allgemeinen und so auch für eine intuitive Benutzung des Systems. Laut Hacker (1986) repräsentiert dieses handlungsleitende mentale Modell die mentalen Repräsentationen des *antizipierten Handlungsziels*, der *Ausführungsbedingungen* und der *Transformationsmaßnahmen* im Arbeitsgedächtnis. Die Richtigkeit und Differenziertheit des, für den Aufbau des Modells abgerufenen, Vorwissens aus dem Langzeitgedächtnis bestimmt dabei generell über das Ausmaß der Effektivität, der Zufriedenstellung und der mentalen Effizienz bei der Anwendung des mentalen Modells während der Handlungsregulation (Frese & Zapf, 1994; Hacker, 1986; Hacker & Sachse, 2013; Nerdinger et al., 2014; Zacher, 2017). Mit der tatsächlichen Umsetzung der Transformationsmaßnahmen (d.h. Antwortgenerator, Effektorsystem) wird letztendlich ein bestimmtes Verhalten sichtbar und die eigentliche Handlung motorisch ausgeführt (Frese & Zapf, 1994; Hacker, 1986; Hacker & Sachse, 2013; Nerdinger et al., 2014; Zacher, 2017; Zempel, 2003).

Der sequentielle Prozess des Aufbaus und der Anwendung eines solchen handlungsleitenden mentalen Modells mithilfe der mnestischen Systeme, wird im Rahmen der Handlungsregulationstheorie anhand von fünf Phasen verdeutlicht, die der Nutzer bei jeder Handlung durchläuft. Dabei ist jedoch die Abfolge dieser Phasen eher als Idealsequenz zu verstehen. In der Realität kann es abhängig vom Nutzungskontext immer wieder zu Rücksprüngen, Wiederholungen oder dem Auslassen bestimmter Phasen (z.B. durch vorzeitigen Handlungsabbruch) kommen (Frese & Zapf, 1994; Hacker, 1986; Nerdinger et al., 2014; Zacher, 2017; Zempel, 2003).

In der *Phase der Zielsetzung* legt der Nutzer sein neues Handlungsziel und die damit verbundene Aufgabe fest, die er mit der Systemnutzung verfolgen möchte. Die genaue Formulierung seines neuen Handlungsziels kann dabei unter anderem bezüglich seiner Spezifität (d.h. detailliert oder grob), seiner Langfristigkeit (d.h. langfristig oder kurzfristig), seiner hierarchischen Abstufung (d.h. Zerlegung eines Handlungsziels in mehrere Teil- und Unterziele) und seiner Vernetzung mit anderen Zielen variieren (Frese & Zapf, 1994; Zacher, 2017; Zempel, 2003). Das neue Handlungsziel (inkl. der damit verbundenen Zielhierarchie) liegt dem Nutzer am Ende dieser Phase als handlungsleitende mentale Repräsentation im Arbeitsgedächtnis vor und fungiert für die gesamte Handlungsregulation (oder künftige Handlungen, siehe Dijksterhuis & Aarts, 2010) als das relativ stabile antizipierte Resultat der Handlung (d.h. Soll-Zustand). Es dient damit als Vergleichsgröße für die über Rückkopplungsprozesse gesteuerte Handlungsregulation (Frese & Zapf, 1994; Hacker, 1986; Hacker & Sachse, 2013; Nerdinger et al., 2014).

Um sein zuvor definiertes Handlungsziel erreichen zu können, ruft der Nutzer in der *Phase der Orientierung* alle relevanten Schemas aus seinem Langzeitgedächtnis auf Basis der eingehenden Informationen aus seiner physischen (z.B. System) und sozialen Umgebung (z.B. Arbeitskollegen) zur Prüfung der aktuellen Ausgangsbedingungen der Zielerreichung (d.h. Ist-Zustand) und der damit verbundenen Handlungsmöglichkeiten (d.h. Transformationsmöglichkeiten zur Erreichung des Soll-Zustands) ab. Dieser Prozess wird auch als Vorwissensaktivierung oder als Anwendung handlungsrelevanten Vorwissens bezeichnet (siehe Krause & Stark, 2006) und leitet die kognitive Informationsverarbeitung zum Aufbau des mentalen Modells des Nutzungskontextes. Die Effektivität der Anwendung hängt dabei individuell vom Vorwissen des Nutzers bezüglich der geplanten Handlung (z.B. Vorerfahrung mit ähnlichen Handlungen, Wissen über Ausführungsmöglichkeiten und Handlungseinschränkungen, Wissen über Verfügbarkeit von Werkzeugen und Methoden) und den Charakteristika des Nutzungskontextes (z.B. Salienz, Verständlichkeit, Kompatibilität und Konsistenz von Umgebungsreizen aus der physischen und sozialen Umgebung) ab (Frese & Zapf, 1994; Hacker, 1986; Hacker & Sachse, 2013; Nerdinger et al., 2014; Zacher, 2017; Zempel, 2003).

In der *Phase der Plangenerierung und Entscheidung* leitet der Nutzer auf Basis seines in der letzten Phase aufgebauten mentalen Modells geeignete Handlungspläne (d.h. zielgerichtete Aktionsfolgen) ab. Diese Handlungspläne fungieren für den Nutzer als adäquate mentale Simulationen (d.h. mentales Probehandeln) der durchzuführenden Aktionen (d.h. Transformationen) zur Erreichung seines Handlungsziels (Hacker & Sachse, 2013; Nerdinger et al., 2014). Dabei können unter anderem der Detaillierungsgrad der Handlungspläne, die Anzahl hierarchischer Abstufungen eines Handlungsplans, die Bewusstheit eines Handlungsplans und die Berücksichtigung von Alternativplänen variieren (Frese & Zapf,

1994; Zacher, 2017; Zempel, 2003). Die Phase der Plangenerierung und Entscheidung endet damit, dass sich der Nutzer für die Durchführung eines bestimmten Handlungsplans entscheidet und damit der Wechsel von der Handlungsvorbereitung zum tatsächlichen Handlungsvollzug erfolgt (Hacker, 1986).

In der *Phase der Handlungsausführung* führt der Nutzer die in der vorigen Phase kognitiv vorbereiteten Aktionsfolgen (z.B. Mausklicks) zur Zielerreichung motorisch durch, die er zuvor bei der mentalen Simulation durchgespielt hat (Frese & Zapf, 1994; Hacker, 1986; Nerdinger et al., 2014; Zacher, 2017; Zempel, 2003).

In der *Phase der Feedbackverarbeitung* verarbeitet der Nutzer letztendlich Feedback, welches ihm signalisiert, ob er sein Handlungsziel durch die Handlungsausführung bereits erreicht hat, er etwas bei seinen Transformationsmaßnahmen (d.h. Handlungsplänen) oder dem Handlungsziel selbst ändern muss (d.h. Anpassung des mentalen Modelles, Änderung des Handlungsziels bzw. der Teilziele, Änderung des Handlungsplans) oder die Handlungsregulation generell beendet (d.h. Zielablösung, siehe Haase, Heckhausen, & Wrosch, 2013) werden sollte (Zacher, 2017). Feedback kann unter anderem bezüglich folgender Merkmale variieren: parallel oder nach der Handlung, intern oder extern, unmittelbar oder verzögert, positiv oder negativ, detailliert oder weniger detailliert (Frese & Zapf, 1994; Zacher, 2017; Zempel, 2003).

Die mit der Handlungsregulation verbundene kognitive Informationsverarbeitung (siehe Abbildung 2.2) kann sich für den Nutzer beim Aufbau seines mentalen Modells, in Abhängigkeit von der gerade durchlaufenden Phase und seines Vorwissens, mehr oder weniger aktiv gestalten. Seine kognitive Informationsverarbeitung kann dementsprechend während des sequentiellen Prozesses auf einem Kontinuum zwischen unbewusster (d.h. *unbewusste Verarbeitung oder Assimilation*: holistische Integration eingehender Informationen in bereits bestehende Schemas, ohne diese zu verändern) und bewusster kognitiver Informationsverarbeitung (d.h. *bewusste Verarbeitung oder Akkommodation*: Erstellung oder Veränderung von Schemas basierend auf eingehenden neuen Informationen) variieren (siehe Piaget, 1976; Saifoulline & Hemberger, 2011; Seel, 1991). Laut Seel (1991) sind beide kognitive Informationsverarbeitungsprozesse dabei komplementär. In der ersten Phase der Handlungsregulation wird durch Akkommodation ein neues Handlungsziel für die bevorstehende Handlung im Arbeitsgedächtnis bewusst festgelegt, wenn nicht ein bereits festgelegtes Ziel verfolgt werden soll. In der zweiten Phase der Handlungsregulation werden, auf Basis des zu verfolgenden Handlungsziels mithilfe des Arbeitsgedächtnisses und dessen Subsystemen, die aus der Umgebung eingehenden handlungsrelevanten Informationen bezüglich der *aktuellen Ausgangsbedingungen der Zielerreichung* (d.h. Ist-Zustand) geprüft.

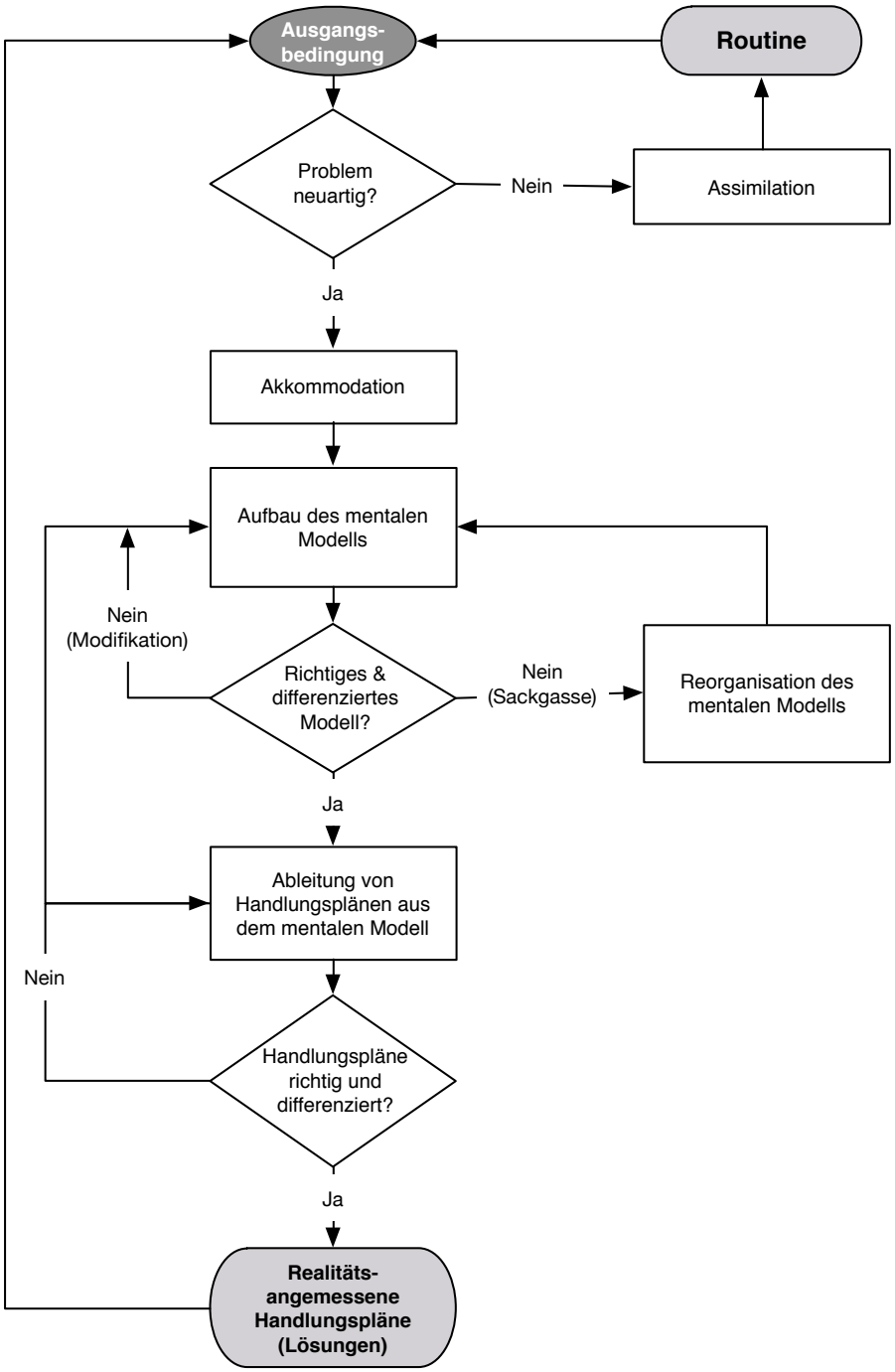


Abbildung 2.2. Flussdiagramm der kognitiven Informationsverarbeitung mittels Akkommodation in Anlehnung an Saifoulline und Hemberger (2011). Die mithilfe von Rauten dargestellten Entscheidungen werden mithilfe von präattentiven metakognitiven Prozessen (siehe Abschnitt 2.2) getroffen.

Beim Prüfen des Ist-Zustands kommen präattentive metakognitive Prozesse im Arbeitsgedächtnis zum Einsatz (siehe Stanovich, West, & Toplak, 2014), die darüber entscheiden, ob auf Basis der vorliegenden aktuellen Ausgangsbedingungen der Zielerreichung (d.h. Ist-Zustand) die Erreichbarkeit des Handlungsziels im vorliegenden Nutzungskontext als nicht neuartig und damit als routinemäßig eingestuft werden kann (siehe Abbildung 2.2). Metakognitive Prozesse können als selbstregulatorische Prozesse der Handlungsregulation verstanden werden, die bereits a priori Hinweise bieten, wie Handlungssituationen analysiert, strukturiert, und verarbeitet werden können (Zempel, 2003).

An dieser Stelle sei zusätzlich darauf hingewiesen, dass der Begriff *Metakognition* in diversen Forschungsarbeiten oftmals auch als eine *bewusste* Auseinandersetzung bzw. Reflektion mit den eigenen kognitiven Prozessen (z.B. Lernen, Gedächtnis, Kreativität) beschrieben wird (Dunlosky & Metcalfe, 2008; Flavell, 1979; Hasselhorn & Artelt, 2018). Insbesondere im Lernprozess spielt eine bewusste Reflektion über eigene Gedanken, Wissen und Verhalten eine wichtige Rolle (Hasselhorn & Artelt, 2018). Im Rahmen dieser Arbeit soll Metakognition jedoch nur in Form von präattentiven und damit überwiegend *unbewussten* metakognitiven Prozessen verstanden werden. Auf die genaue Funktionsweise derartiger Prozesse wird erst in Abschnitt 2.2 vertieft eingegangen, da metakognitive Aspekte generell nicht von der IUUI-Forschergruppe in ihren Arbeiten thematisiert werden und ferner auch nicht ausgiebig mithilfe der Handlungsregulationstheorie von Hacker (1986) erklärbar sind.

Kann die Handlungssituation auf Basis des Ist-Zustands mithilfe präattentiver metakognitiver Prozesse als nicht neuartig eingestuft werden (siehe „Problem neuartig?“ in Abbildung 2.2), so startet auf der kognitiven Ebene der implizite Informationsverarbeitungsprozess der Assimilation (siehe Piaget, 1976; Saifoulline & Hemberger, 2011; Seel, 1991), wie es auch Abbildung 2.2 zeigt. Hier aktivieren die im Kurzzeitgedächtnis repräsentierten Merkmale des Ist-Zustands unbewusst vorhandene Schemas aus dem Langzeitgedächtnis. Die aktivierten Schemas verfügen über sog. Leerstellen (d.h. Slots, offene Umgebungsparameter), die bei der Assimilation mit Informationen aus der Umwelt gefüllt und die Schemas so flexibel an die Umgebung angepasst werden können (Anderson & Funke, 2001; Kopp & Mandl, 2005; Rumelhart, 2017). Bei diesem unbewussten Prozess werden Schemas durch die Umgebungsinformationen nur geringfügig verändert. Das handlungsleitende mentale Modell steht dem Nutzer unmittelbar zur Verfügung und konkrete Handlungspläne für die aktuelle Handlungssituation können in der folgenden Phase der Plangenerierung abgeleitet werden (Saifoulline & Hemberger, 2011; Seel, 1991).

Wird bei der Prüfung des Ist-Zustands mithilfe präattentiver metakognitiver Prozesse in der Phase der Orientierung jedoch festgestellt, dass die Handlungssituation stattdessen als neuartig einzustufen ist (siehe „Problem neuartig?“ in Abbildung 2.2), sind bereits vorhandene Schemas für eine Assimilation nicht geeignet und ein darauf aufbauendes mentales Modell für die aktuelle Handlungssituation nicht richtig und differenziert genug. Deswegen muss auf kognitiver Ebene zunächst ein für die Handlung angemessenes mentales Modell durch den expliziten Prozess der Akkommodation (siehe „Aufbau des mentalen Modells“ in Abbildung 2.2) aufgebaut werden (Saifoulline & Hemberger, 2011; Seel, 1991). Hierzu müssen alle eingehenden Umgebungsinformationen zunächst im Arbeitsgedächtnis repräsentiert werden (d.h. Assimilation hat die gleiche Grundvoraussetzung wie Akkommodation, siehe Seel, 1991), um diese mit im Langzeitgedächtnis enthaltenen Schemas in der Phase

der Orientierung abzugleichen und so nach Analogien suchen zu können. Um Analogien zu finden, benötigt es eine Reihe von bewussten Filter- und Analyseprozessen der zentralen Exekutive, auf Basis derer exekutiver Grundfunktionen zur Bestimmung der Eignung der im Langzeitgedächtnis vorhandenen Schemas für die aktuelle Handlungssituation. Wurde eine passende Analogie gefunden, fungiert diese als Grundlage zur Lösung der Probleme der neuartigen Situation (Saifoulline & Hemberger, 2011).

In einem nächsten Schritt wird das neu erstellte mentale Modell auf seine Richtigkeit und Differenziertheit zur Erreichung des Handlungsziels geprüft (siehe „Richtiges und differenziertes Modell?“ in Abbildung 2.2). Falls das mentale Modell nicht als richtig und differenziert eingestuft wird (d.h. Handlungssituation mit dem System wird durch das Modell nicht realitätsangemessen repräsentiert), kommt es zur weiteren Akkommodation (siehe „Nein (Modifikation)“ in Abbildung 2.2), bis das mentale Modell im besten Falle als hinreichend richtig und differenziert angesehen wird. Mithilfe der zentralen Exekutive sind aus diesem Modell Handlungspläne in der nächsten Phase der Handlungsregulation ableitbar (siehe „Ableitung von Handlungsplänen aus dem mentalen Modell“ in Abbildung 2.2). Wenn diese Akkommodationsprozesse fehlschlagen und das Modell auf Basis metakognitiver Prozesse weiterhin als nicht richtig und differenziert genug eingestuft wird (siehe „Nein (Sackgasse)“ in Abbildung 2.2), ist eine noch umfassendere Umstrukturierung des mentalen Modells erforderlich und es erfolgt ein Rücksprung zur Phase der Orientierung (siehe „Reorganisation des mentalen Modells“ in Abbildung 2.2). Hierbei werden grundlegende Variablen des mentalen Modells neu definiert bzw. ausgetauscht. Auf diese Weise ändert der Nutzer die Sichtweise auf sein Handlungsproblem grundlegend und nimmt beispielsweise andere kausale Bedingungen zwischen den Variablen an, weswegen der Regulationsprozess erneut durchlaufen wird (Saifoulline & Hemberger, 2011).

Schließlich sollten die aus einem richtigen und differenzierten handlungsleitenden mentalen Modell abgeleiteten Handlungspläne ebenfalls selbst auf ihre Eignung zur Erreichung des Handlungsziels hin geprüft werden (siehe „Handlungspläne richtig und differenziert?“ in Abbildung 2.2). Fällt diese Prüfung negativ aus, folgt ein erneuter Prozess zum Aufbau des mentalen Modells bzw. zur Ableitung von Handlungsplänen. Ist die Prüfung positiv, werden die Handlungspläne als zur Problemlösung geeignet angesehen (siehe „Realitätsangemessene Handlungspläne (Lösungen)“ in Abbildung 2.2). Infolgedessen kann die daraufhin gestartete Handlungsdurchführung die Ausgangsbedingung verändern (Saifoulline & Hemberger, 2011). Die Lösung des Handlungsproblems mittels unbewusster kognitiver Assimilationsprozesse ist für den Nutzer in der Regel effektiver, zufriedenstellender und mental effizienter als mittels bewusster kognitiver Akkommodationsprozesse, da bei der Assimilation lediglich bereits bestehende Schemas unbewusst aktiviert und Kontextinformationen ohne Anpassung der Schemas holistisch integriert werden müssen (siehe Kopp & Mandl, 2005; Saifoulline & Hemberger, 2011; Seel, 1991).

Jedoch kann es nicht nur zu Rücksprüngen zwischen den einzelnen Phasen der Handlungsregulation innerhalb einer Verarbeitungsart kommen, wie es die Abbildung 2.2 bereits anhand der Akkommodation veranschaulichte, sondern es kann auch zu dynamischen Wechseln zwischen den beiden archetypischen Verarbeitungsmodi führen. So kann es bei der Assimilation zu einer unvorhergesehenen Situation (z.B. ausgewählter Handlungsplan funktioniert nicht) kommen, in der die aktivierten Schemas für eine implizite kognitive Verarbeitung nicht mehr ausreichen. Dieser Zustand wird dem Handelnden durch einen

unbewussten affektiven metakognitiven Richtimpuls (d.h. Fringe, siehe Price & Norman, 2008) signalisiert, und er kann sein Modell mithilfe von Akkommodation an diese Situation anpassen (siehe Mangan, 2015; E. Norman, 2017; E. Norman, Price, & Duff, 2010; R. Reber et al., 2004; Schwarz, 2015; Stanovich et al., 2014; V. Thompson, Turner, & Pennycook, 2011; Topolinski, 2011; Topolinski & Strack, 2009; Zander, Öllinger, & Volz, 2016).

Umgekehrt kann es bei der bewussten kognitiven Verarbeitung (d.h. Akkommodation) während des Aufbaus des mentalen Modells auch zu einem derartigen Richtimpuls kommen und passieren, dass das durch Akkommodation aufgebaute mentale Modell durch erhöhte Übung (d.h. Intensität und Häufigkeit der Anwendung, siehe Bargh & Chartrand, 1999) richtig und differenziert genug ist, sodass damit eine unbewusste Assimilation möglich und deswegen keine Akkommodation mehr nötig ist (siehe Mangan, 2015; E. Norman, 2017; E. Norman et al., 2010; R. Reber et al., 2004; Stanovich et al., 2014; V. Thompson et al., 2011; Topolinski, 2011; Topolinski & Strack, 2009; Zander et al., 2016). Da der Nutzer im Zuge seiner Systemnutzung auf mehr oder weniger bekannte Probleme stößt (d.h. Handlungsziele sind oft hierarchisch organisiert und nicht alle Teilziele sind gleich komplex bzw. neuartig, siehe Volpert, 1982) und die Richtigkeit, sowie Differenziertheit seines Vorwissens bezüglich dieser Probleme wiederum variieren kann (d.h. Vorwissen setzt sich auf Basis des Kontinuums der Wissensquellen aus verschiedenen Quellen zusammen, die sich nicht gleichermaßen alle gut für alle Handlungssituationen eignen, siehe Hurtienne, 2011), muss er dynamisch zwischen unbewusster und bewusster Verarbeitung für eine effektive Handlungsregulation mithilfe seiner exekutiven Grundfunktion, dem Shifting, hin und her wechseln können (A. Miyake et al., 2000). Demzufolge kann eine Handlung nicht per se mittels reiner unbewusster oder bewusster kognitiver Informationsverarbeitung reguliert werden (Nerdinger et al., 2014).

Laut O'Brien et al. (2008) bildet genau die Kombination von unbewusster (d.h. Assimilation) und bewusster Informationsverarbeitung (d.h. Akkommodation) die Basis für menschliches Denken, sowie für zielgerichtetes Handeln. Es kann deswegen neben diesen beiden archetypischen Modi der kognitiven Informationsverarbeitung auch ein dritter Modus beschrieben werden, der sich auf einem Kontinuum zwischen diesen beiden Extremen befindet. Kognitive Informationsverarbeitungsprozesse, die kontinuierlich zwischen bewusster und unbewusster kognitiver Informationsverarbeitung dynamisch wechseln können, werden oft auch als bewusstseinsfähige Prozesse bezeichnet (Frese & Zapf, 1994; Hacker & Sachse, 2013; Hurtienne, 2011). Eine detaillierte Erläuterung des Zusammenspiels von unbewusster Assimilation und bewusster Akkommodation unter Berücksichtigung von metakognitiven Kontroll-/Steuerungsprozessen (d.h. präattentive metakognitive Prozesse) bei der Vorwissensaktivierung wird im nächsten Abschnitt anhand von Zwei-Prozess-Theorien diskutiert (siehe Abschnitt 2.2), da die Handlungsregulationstheorie (siehe Hacker, 1986) und assoziierte Informationsverarbeitungsmodelle (siehe Wickens et al., 2015) als ganzheitliche Modelle eher den Anspruch erheben, das menschliche Handeln und Problemlösen im Allgemeinen zu erklären und sich dementsprechend bezüglich der detaillierten Funktionsweise des Wechsels, zwischen unbewusster und bewusster kognitiver Informationsverarbeitung, nur indirekt äußern (Betsch et al., 2011; Nerdinger et al., 2014).

Für eine Erklärung, wie man anhand des Vorhandenseins von unbewusster und bewusster kognitiver Informationsverarbeitung feststellen kann, ob es sich gerade um eine intuitive

oder eine nicht intuitive Handlung handelt, reicht die alleinige Betrachtung der Handlungsregulation aus der vorgestellten sequentiellen Phasenperspektive nicht aus. Stattdessen muss die Handlungsregulation zusätzlich aus einer hierarchischen Ebenenperspektive betrachtet werden. Bei dieser hierarchischen Ebenenperspektive (siehe Abbildung 2.3), wird der für die Anwendung des handlungsrelevanten Vorwissens benötigte Bewusstseinsgrad durch drei verschiedene kognitive Informationsverarbeitungsebenen (d.h. intellektuelle Ebene, perzeptiv-begriffliche Ebene, sensomotorische Ebene) vereinfachend veranschaulicht (Frese & Zapf, 1994; Hacker, 1986; Hacker & Sachse, 2013; Nerdinger et al., 2014). Frese und Zapf (1994) unterscheiden im Vergleich zur Originalarbeit von Hacker (1986) zusätzlich noch eine vierte Ebene des abstrakten Denkens, auf der die bereits kurz angesprochenen präattentiven metakognitiven Prozesse angesiedelt sind, die für den dynamischen Wechsel zwischen den Ebenen zuständig sind (siehe Abschnitt 2.2.2). Eine solche vierte Ebene wird in der Handlungsregulationstheorie aufgrund der dadurch erhöhten Komplexität nicht durchgängig verwendet und präattentive metakognitive Aspekte werden oft nicht thematisiert (siehe Hacker & Sachse, 2013; Nerdinger et al., 2014).

Die sehr vereinfachte Erklärung der kognitiven Informationsverarbeitung mithilfe von drei Ebenen (siehe Abbildung 2.3) ohne expliziten Verweis auf eine übergreifende, regulierende, präattentive, überwiegend unbewusste Metakognition ist in der HCI am weitesten verbreitet (z.B. Nielsen, 1994; Wickens et al., 2015; Zandbergen, 2015; Zapf, Brodbeck, & Prümper, 1989). Die genaue Funktionsweise des kontinuierlichen Wechsels zwischen unbewusster und bewusster Verarbeitung wird bei der Verwendung von drei Ebenen nicht thematisiert und stattdessen eine bewusstseinsfähige „Zwischenebene“ eingeführt (Semmer & Pfäfflin, 1978), um die Möglichkeit eines dynamischen Wechsels zu symbolisieren (siehe Zempel, 2003). Die HCI nutzt als theoretische Grundlage hauptsächlich die beschriebene Handlungsregulationstheorie (siehe Hacker, 1986), das damit fast identische Drei-Ebenen-Modell (siehe Rasmussen, 1983) und weiterführende Informationsverarbeitungsmodelle wie das Modell der Informationsverarbeitung von Rasmussen (1986) oder das Modell von D. A. Norman und Shallice (1986).

Auf die letzteren beiden Informationsverarbeitungsmodelle stützt auch die IUUI-Forschergruppe (siehe Hurtienne, 2011) ihre Argumentation und veranschaulicht den Zusammenhang zwischen intuitiver Benutzung und dem Bewusstseinsgrad bei der Anwendung von Vorwissen anhand von drei verschiedenen kognitiven Informationsverarbeitungsebenen (d.h. bewusstseinsunfähig/unbewusst, bewusstseinsfähig/teil-unbewusst und bewusstseinspflichtig/bewusst). D. A. Norman und Shallice (1986) bezeichnen die bewusstseinsunfähige Ebene als „vollständig automatisch“, die bewusstseinsfähige Ebene als „teilweise automatisch“ und die bewusstseinspflichtige Ebene als „willentliche Steuerung“, wobei der Begriff der Automatik für eine eindeutige Unterscheidung der Ebenen sehr vage ist. Da auch Rasmussen (1983) für die Unterscheidungen seiner Informationsverarbeitungsebenen die ebenfalls sehr vagen Begriffe „wissensbasiert“, „regelbasiert“ und auf „sensorischen Fähigkeiten basiert“ verwendet und die Wahl dieser Begriffe zusätzlich den Eindruck erwecken könnte, dass lediglich auf der erstgenannten Ebene handlungsrelevantes Vorwissen als Grundlage für den Aufbau des mentalen Modells genutzt wird (Hacker, 2009), soll sich zur Wahrung von Klarheit im Rahmen dieser Arbeit an den Begriffen aus der Handlungsregulationstheorie von Hacker (1986) orientiert werden.

Die drei im Rahmen der Handlungsregulation angenommenen kognitiven Informationsverarbeitungsebenen (siehe Abbildung 2.3) sind laut Hacker und Sachse (2013) hierarchisch bzw. heterarchisch „verschachtelt“, weswegen untergeordnete Ebenen ihre Resultate in Form von Rückkopplungen an die übergeordneten Ebenen weiterleiten, um somit gegebenenfalls korrigierend eingreifen zu können. Auf diese Weise können die eigentlich von metakognitiven Prozessen übernommenen Steuerungs- und Kontrollfunktionen von den drei Ebenen theoretisch selbst übernommen werden, weswegen eine vierte Ebene nicht zwingenderweise nötig ist. Höhere Ebenen determinieren jedoch nicht vollkommen über die Regulation auf den unteren Ebenen, denn das Feedback von niedrigeren Ebenen kann auch zu Änderungen auf oberen Ebenen (z.B. Zieländerung des Nutzers, wenn der Cursor nicht mehr reagiert) führen (Frese & Zapf, 1994; Hacker, 2005; Hacker & Sachse, 2013; Nerdinger et al., 2014; Zacher, 2017; Zacher & Frese, 2018). Aus einer Handlungsperspektive sind diese Hierarchieebenen dementsprechend relativ schwach und erlauben einen dynamischen Wechsel zwischen den Regulationsebenen abhängig vom Ausmaß der Richtigkeit und Differenziertheit des mentalen Modells für den jeweiligen Nutzungskontext, so wie dieser bereits bei den Erläuterung zu Abbildung 2.2 ausgiebig besprochen wurde.

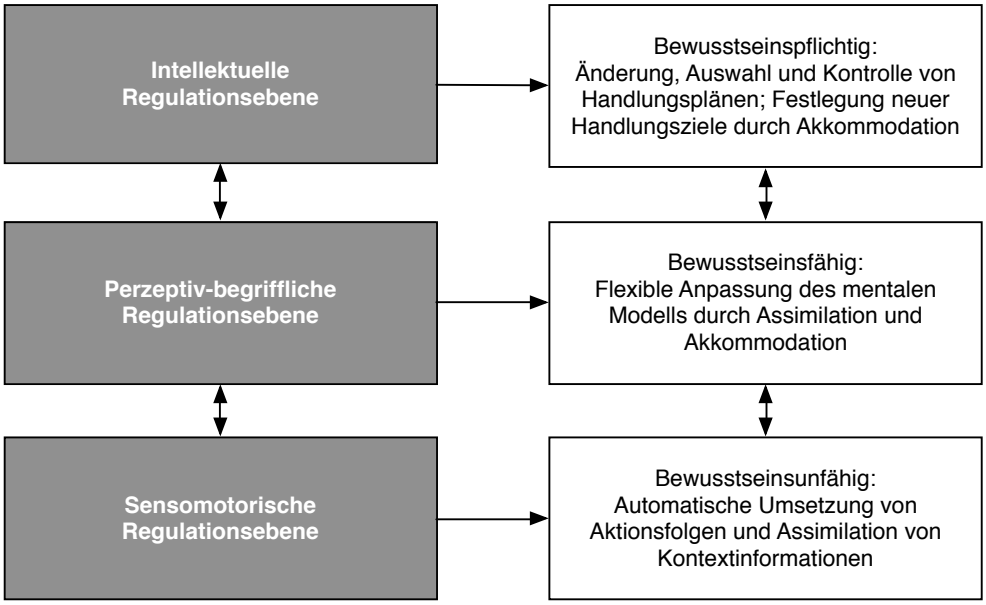


Abbildung 2.3. Zusammenhang von Regulationsebenen und Funktionen in Anlehnung an Hacker (1986) und Nerdinger, Blickle, Schaper und Schaper (2014) unter Berücksichtigung der Terminologie von Seel (1991) und Saifoulline und Hemberger (2011).

Auf der niedrigsten Ebene (siehe Abbildung 2.3), der *sensomotorischen Ebene*, erfolgt die konkrete Ausführung von Handlungsplänen und damit die automatische Umsetzung von motorischen Aktionsfolgen (Frese & Zapf, 1994; Hacker & Sachse, 2013; Nerdinger et al., 2014; Zacher & Frese, 2018; Zempel, 2003), sowie die Assimilation von Kontextinformationen durch unbewusst aktivierte Schemas (siehe Saifoulline & Hemberger, 2011; Seel, 1991). Regulationsvorgänge auf dieser Ebene werden von höheren Ebenen mithilfe präattentiver metakognitiver Prozesse angestoßen (z.B. Schemas können aufgrund einer bekannten Ausgangssituation unbewusst aktiviert und durch Assimilation verarbeitet wer-

den) oder unterbrochen (z.B. Schemas reichen für eine unbekannte Situation nicht mehr aus und das mentale Modell muss durch Akkommodation optimiert werden) (siehe Saifoulline & Hemberger, 2011; Seel, 1991). Die damit mit dieser Ebene in Zusammenhang gebrachte mentale Beanspruchung ist im Vergleich zu den anderen zwei Ebenen gering bzw. die mentale Effizienz hoch, und ist somit auf dieser Ebene zu vernachlässigen (Frese & Zapf, 1994; Nerdinger et al., 2014; Zempel, 2003). Beim Feedback, das im Zuge der Handlungsregulation auf dieser Ebene für höhere Ebenen erzeugt und dort „erlebt“ wird, handelt es sich um die automatische Generierung von kinästhetischen und propriozeptiven Signalen zur Steuerung motorischer Bewegungen (Zacher, 2017).

Dieser Regulationsebene übergeordnet ist die *perzeptiv-begriffliche Regulationsebene* (siehe Abbildung 2.3), die auch als Ebene der flexiblen Handlungsmuster bezeichnet wird (Frese & Zapf, 1994; Hacker & Sachse, 2013; Nerdinger et al., 2014; Zacher, 2017; Zacher & Frese, 2018; Zempel, 2003). Wie bereits erwähnt, handelt es sich bei dieser Ebene laut Semmer und Pfäfflin (1978) um eine „Zwischenebene“ zwischen unbewusster sensomotorischer und bewusster, intellektueller Regulation. Wie im Rahmen der Erläuterungen zur Abbildung 2.2 erwähnt, kann es auch bei routinierten Handlungen, die eigentlich per Assimilation reguliert werden können, zu unvorhergesehenen Ereignissen kommen, die eine dynamische Anpassung des mentalen Modells mittels Akkommodation nötig machen. Umgekehrt können neuartige Handlungen, die eigentlich per Akkommodation reguliert werden, durch viel Übung ab einem gewissen Zeitpunkt mittels Assimilation reguliert werden (Saifoulline & Hemberger, 2011; Seel, 1991).

Es finden auf dieser Ebene entsprechend die damit verbundenen flexiblen Anpassungen des mentalen Modells (siehe Abbildung 2.2) statt, die mit einer geringeren mentalen Beanspruchung und damit höherer mentaler Effizienz einhergehen, als bei der höheren intellektuellen Regulationsebene, wobei die mentale Beanspruchung jedoch höher als die auf der sensomotorischen Ebene ist (Frese & Zapf, 1994; Nerdinger et al., 2014; Zempel, 2003). Wie bereits angesprochen, wird anhand wahrnehmungsinterner metakognitiver Prozesse signalisiert (siehe Stanovich et al., 2014), inwiefern eine Assimilation oder Akkommodation für die Bereitstellung eines richtigen und differenzierten mentalen Modells für die aktuelle Handlungssituation benötigt wird (siehe Saifoulline & Hemberger, 2011; Seel, 1991). In der vereinfachten Drei-Ebenen-Darstellung wird diese Aufgabe von der perzeptiv-begrifflichen Ebene übernommen (siehe Hacker, 1986; Semmer & Pfäfflin, 1978). Auf Basis dieser präattentiven metakognitiven Prozesse können die anderen beiden Ebenen angestoßen und die Verarbeitung auf diese Ebenen verlagert werden. Assimilationsprozesse für die Integration in das bereits vorhandene Modell können somit von dieser Ebene unmittelbar nach unten angestoßen werden (d.h. Delegation an sensomotorische Ebene) und Akkommodationsprozesse für die Anpassung des Modells werden durch eine Rückkopplung an die oberste Ebene durch die Unterstützung metakognitiver Prozesse veranlasst. Auf diese Weise kann das mentale Modell an unterschiedliche Handlungssituationen dynamisch angepasst werden (siehe Frese & Zapf, 1994; Hacker & Sachse, 2013; Zacher, 2017; Zempel, 2003).

Auf der obersten Ebene, der *intellektuellen Ebene* (siehe Abbildung 2.3), findet der Entwurf, die Änderung, die Auswahl und die Kontrolle von Handlungsplänen für neuartige Handlungen, sowie die Festsetzung eines neuen Handlungsziels durch Akkommodation

statt (Frese & Zapf, 1994; Hacker & Sachse, 2013; Nerdinger et al., 2014; Zacher, 2017; Zacher & Frese, 2018; Zempel, 2003). Darüber hinaus wird auf dieser Ebene auch der Aufbau eines richtigen und differenzierten mentalen Modells durch Akkommodation durchgeführt (siehe Saifoulline & Hemberger, 2011; Seel, 1991) und die flexiblen Anpassungen dieses mentalen Modells auf der perzeptiv-begrifflichen Ebene mithilfe von präattentiven metakognitiven Prozessen angestoßen. Während die Handlungsregulation auf dieser Ebene eine hohe mentale Beanspruchung und damit eine geringe mentale Effizienz im Vergleich zu den anderen beiden Ebenen mit sich bringt (Frese & Zapf, 1994; Nerdinger et al., 2014; Zempel, 2003), kann durch Akkommodation ein neues mentales Modell richtig und differenziert genug aufgebaut werden (siehe Saifoulline & Hemberger, 2011; Seel, 1991), sodass eine Handlungsregulation auf den unteren Ebenen möglich ist (siehe Frese & Zapf, 1994; Hacker & Sachse, 2013; Zacher, 2017).

Da eine Handlung laut der Definition der IUUI-Forschergruppe immer mental effizient sein muss, um als intuitiv gelten zu können, erfüllen nur die unteren beiden kognitiven Informationsverarbeitungsebenen diese Anforderung (Hurtienne, 2011). Die unbewusste Anwendung von Vorwissen als Grundvoraussetzung intuitiver Benutzung lässt sich dementsprechend als eine mental effiziente und dabei gleichzeitig effektive Anwendung von (handlungsrelevantem) Vorwissen operationalisieren, die somit entweder vollständig unbewusst (d.h. bewusstseinsunfähig, automatisch, sensomotorisch) oder teil-unbewusst (d.h. bewusstseinsfähig, teil-automatisch, perzeptiv-begrifflich) abläuft (siehe Hurtienne, 2011). Eine vollständig bewusste kognitive Verarbeitung auf der oberen Ebene wird damit von der IUUI-Forschergruppe unter Berücksichtigung der vorgestellten hierarchischen Ebenenperspektive explizit ausgeschlossen, da hier aufgrund der mangelnden mentalen Effizienz (d.h. hohe mentale Beanspruchung) überhaupt keine intuitive Benutzung vorliegen kann (Hurtienne, 2011; Naumann et al., 2007). Intuitive Benutzung ist jedoch aufgrund des zielgerichteten Informationsaustausches zwischen System und Nutzer immer als eine Handlung zu betrachten, die per Definition zumindest theoretisch auch bei einer vollständigen intuitiven Handlungsregulation immer alle drei Ebenen einschließen kann (Hacker, 1986, 2009; Hacker & Sachse, 2013). Erfolgt die anfängliche Zielformulierung (d.h. „Was will ich mit dem System machen?“) beispielsweise bewusst durch Akkommodation (d.h. Festlegung eines neuen Handlungsziels), wie es auch im Rahmen der Beschreibung von Abbildung 2.2 erklärt wurde, so findet diese Neuformulierung auf der oberen Ebene statt und stößt die Regulationen auf den unteren Ebenen an (Hacker, 1986).

Laut aktueller Forschung kann es jedoch auch vorkommen, dass keine explizite, bewusste Zielformulierung per Akkommodation stattfindet und bereits existierende Ziele für die Handlungsregulation übernommen werden, da Ziele ebenfalls als mentale Repräsentationen im handlungsleitenden mentalen Modell vorliegen. Diese umfassen zusätzlich noch die Ausführungsbedingungen und die Transformationsmaßnahmen (d.h. Handlungspläne) bezüglich des jeweiligen Handlungsziels im Arbeitsgedächtnis (Bargh & Gollwitzer, 1994; R. Cooper & Shallice, 2006; Dijksterhuis & Aarts, 2010; Hacker, 1986). Derartige Forschung kann somit als eine Weiterentwicklung gegenüber der klassischen Handlungsregulationstheorie aufgefasst werden. Menschen können dementsprechend auch die bereits im mentalen Modell hinterlegten Ziele unbewusst verfolgen ohne diese nochmal bewusst formulieren zu müssen (d.h. Ziele, die irgendwann mal bewusst per Akkommodation formuliert wurden und dementsprechend keine Neuformulierung eines Ziels per Akkommodation erforderlich

ist), wenn die damit verbundenen mentalen Zielrepräsentationen durch Verhaltens- und Kontextinformationen unbewusst angestoßen werden (Dijksterhuis & Aarts, 2010).

Beim abschließenden Soll-Ist-Vergleich (d.h. „Habe ich mit dem System mein Ziel bereits erreicht?“) verhält es sich, basierend auf aktueller Forschung, ähnlich. Dieser Soll-Ist-Vergleich kann sowohl bewusst auf der oberen Ebene erfolgen, wie von der klassischen Handlungsregulationstheorie vorgeschlagen (siehe Frese & Zapf, 1994; Hacker, 1986, 2009; Hacker & Sachse, 2013), als auch aufgrund der Tatsache, dass ein bekanntes Ziel ja bereits als mentale Repräsentation aufgrund vorgelagerter Akkommodation vorliegen kann, unbewusst reguliert werden (Bongers, Dijksterhuis, & Spears, 2010; Chartrand & Bargh, 1996; Dijksterhuis & Aarts, 2010; Fishbach, Friedman, & Kruglanski, 2003). Demzufolge ist eine gewisse bewusste Informationsverarbeitung auch bei einer theoretisch vollkommen intuitiven Handlung möglich (siehe Frese & Zapf, 1994; Hacker, 1986, 2009), was eine Regulation auf allen drei Ebenen erfordert. Auf der anderen Seite erscheint es auch bei einer vollkommen bewussten Informationsverarbeitung, mindestens die motorische Umsetzung unbewusst abzulaufen, weswegen hier auch eine gewisse unbewusste Regulation auftreten sollte (Hacker, 2009). Im Allgemeinen besitzt die hierarchische Drei-Ebenen-Perspektive als Vereinfachung zusätzlich das Problem, dass diese, sofern keine vierte metakognitive Ebene berücksichtigt wird, keine Aussagen zur genauen Funktionsweise des Wechsels zwischen den Ebenen anhand metakognitiver Richtimpulse trifft. Diese Aufgabe wird stattdessen auf der mittleren Ebene selbst verortet (Hacker & Sachse, 2013; Nerdinger et al., 2014), weswegen durch einen expliziten Ausschluss der metakognitiven Ebene diese Vereinfachung des gegenseitigen Anstoßens nicht mehr funktioniert.

Demzufolge erscheint unter Berücksichtigung der vereinfachten Darstellung der kognitiven Informationsverarbeitung im Rahmen der Handlungsregulationstheorie, das Merkmal „unbewusste Anwendung von Vorwissen“, welches von der IUUI-Forschergruppe in ihrer Definition als die Grundvoraussetzung intuitiver Benutzung angesehen wird, als unzureichend spezifiziert. Aufgrund des Begriffs „unbewusst“ könnte der Eindruck entstehen, dass explizit die obere Ebene bei intuitiver Benutzung ausgeschlossen wird, was aus einer Handlungsperspektive nicht möglich ist, da die Definition und die Überprüfung des Handlungsziels auf der oberen Ebene stattfinden kann. Darüber hinaus werden in diesem Begriff präattentive metakognitive Prozesse nicht eingeschlossen, welche für einen dynamischen Wechsel zwischen den Ebenen benötigt werden. In Übereinstimmung mit der IUUI-Forschergruppe findet intuitive Benutzung zwar hauptsächlich auf den beiden unteren Ebenen und somit *überwiegend unbewusst* statt, aber es werden trotzdem zusätzlich präattentive metakognitive Prozesse für die Steuerung des Wechsels und die dritte Ebene theoretisch für einige Handlungsphasen auch im Rahmen einer vollkommenen intuitiven Handlung benötigt (siehe Frese & Zapf, 1994; Hacker & Sachse, 2013; Nerdinger et al., 2014).

Dieser bei einer intuitiven Benutzung eingenommene überwiegend unbewusste Bewusstseinsbereich soll im Rahmen dieser Arbeit deswegen nicht mithilfe des Begriffs „unbewusst“ umschrieben werden, sondern stattdessen, wegen der Grenze zum menschlichen Bewusstsein, als *Fringe Consciousness* bezeichnet werden, so wie er auch von zahlreiche anderen Forschern mit Intuition außerhalb der HCI bezeichnet wird (z.B. Mangan, 2015; E. Norman, 2017; E. Norman et al., 2010; Price & Norman, 2008; R. Reber et

al., 2004; Zander et al., 2016). Da sich dieser Bereich schlecht anhand der vereinfachten Drei-Ebenen-Darstellung beschreiben lässt, mit der kein wirkliches Kontinuum dargestellt werden kann, wird im Rahmen dieser Arbeit intuitive Benutzung nicht per se als bewusst, teil-unbewusst oder unbewusst anhand der drei statischen Ebenen charakterisiert. Stattdessen wird die unbewusste Anwendung von handlungsrelevantem Vorwissen als Grundvoraussetzung intuitiver Benutzung als *überwiegend unbewusster kognitiver Informationsverarbeitungsprozess auf Basis handlungsrelevanten Vorwissens an der Grenze zum menschlichen Bewusstsein* angesehen, um auf diese Weise präattentive metakognitive Richtimpulse und auch bewusste Zielformulierungen berücksichtigen zu können. Auf diesen Bereich der Fringe Consciousness wird in Abschnitt 2.2 im Rahmen präattentiver metakognitiver Prozesse genauer eingegangen werden, da die IUUI-Forschergruppe diesen Bereich nicht thematisiert und die Handlungsregulationstheorie auch nur am Rande darauf eingeht (siehe Frese & Zapf, 1994; Hacker & Sachse, 2013; Nerdinger et al., 2014; Zacher & Frese, 2018).

2.1.1.3 Merkmale aus Sicht der IUUI-Forschergruppe

Zusammenfassend lassen sich unter Berücksichtigung der in diesem Abschnitt vorgestellten Literatur folgende charakteristische Merkmale aus der Definition intuitiver Benutzung der IUUI-Forschergruppe ableiten. Die überwiegend unbewusste kognitive Informationsverarbeitung auf Basis des mentalen Modells (d.h. handlungsrelevantem Vorwissen) bildet dabei die Vorbedingung (•) und die anderen Merkmale die objektiven (→) und die subjektiven (⇢) Konsequenzen intuitiver Benutzung. Laut der IUUI-Forschergruppe treten diese Merkmale bei einer intuitiven Benutzung gleichzeitig im gleichen Ausmaß auf bzw. das Ausmaß intuitiver Benutzung spiegelt sich in diesen Merkmalen gleichermaßen wider (Hurtienne, 2011). Die Merkmale wurden so formuliert, sodass sie die vollkommene intuitive Benutzung und daher das positive Extremum des Kontinuums repräsentieren. Je weiter man sich von den aufgelisteten Ausprägungen der Merkmale entfernt, umso geringer wird die intuitive Benutzung des Systems (z.B. weniger mental effizient, weniger effektiv und weniger zufriedenstellend).

Charakteristische Merkmale intuitiver Benutzung aus Perspektive der IUUI-Forschergruppe:

- • Überwiegend unbewusster kognitiver Informationsverarbeitungsprozess auf Basis von handlungsrelevantem Vorwissen an der Grenze zum menschlichen Bewusstsein
- → Mental effiziente kognitive Informationsverarbeitung
- → Effektive Benutzung
- ⇢ Zufriedenstellende Benutzung
  - Wahrgenommene geringe mentale Beanspruchung
  - Wahrgenommene hohe Zielerreichung
  - Wahrgenommene hohe Vertrautheit
  - Wahrgenommener geringer Lernaufwand
  - Wahrgenommene geringe Fehlerrate

2.1.2 QUT-Forschergruppe

Die Definition der australischen Forschungsgruppe um Alethea Blackler der Queensland University of Technology (QUT) besitzt im Vergleich zur IUUI-Definition eine andere theoretische Grundlage. Sie stützt ihre Definition auf eine ausgiebige Literaturrecherche in den Bereichen der intuitiven Entscheidungsfindung, des kreativen Denkens, der Philosophie, der Lernpsychologie und der Neuro- bzw. Kognitionswissenschaft (Blackler, 2006, 2018; Blackler & Popovic, 2015; Blackler, Popovic, & Mahar, 2010; Blackler, Popovic, & Mahar, 2002, 2003). Daraus leitete die Forschergruppe nach mehreren Revisionen (siehe Blackler, 2006) folgende Definition ab:

> Intuitive Benutzung von Produkten passiert auf der Basis von bereits vorhandenem Wissen und Erfahrungen. Sie erfolgt schnell und generell unbewusst, was dazu führt, dass Menschen ihre im Rahmen einer intuitiven Benutzung getroffenen Entscheidungen im Nachhinein nicht erklären können (Blackler, 2006).

Aus der QUT-Definition können die folgenden drei Merkmale intuitiver Benutzung abgeleitet werden: die (1) *unbewusste Anwendung von Vorwissen* während der Benutzung, die (2) *zeitlich effiziente Benutzung* und die (3) *nicht vorhandene retrospektive Verbalisierbarkeit* der Benutzung, welche bei einer intuitiven Benutzung *alle gemeinsam* auftreten sollten. Die QUT-Forschergruppe trennt in ihren Arbeiten nicht zwischen Vorbedingung und Konsequenzen intuitiver Benutzung (siehe Blackler, 2006, 2018; Blackler & Popovic, 2015; Blackler et al., 2010; Blackler et al., 2002, 2003), wobei die unbewusste Anwendung von handlungsrelevantem Vorwissen in Anlehnung an die IUUI-Forschergruppe (siehe Teilabschnitt 2.1.1) als Grundvoraussetzung und die anderen beiden Merkmale als Konsequenzen intuitiver Benutzung interpretierbar sind. Obwohl die QUT-Forschergruppe es nicht explizit in ihrer Definition erwähnt, können alle genannten charakterisierenden Merkmale als Variablen und somit auch das gesamte Ausmaß an intuitiver Benutzung selbst bezüglich der Stärke ihrer Ausprägung auf einem Kontinuum variieren (Blackler, 2006). Dabei treten bei einer intuitiven Benutzung alle Merkmale gleichzeitig und im gleichen Ausmaß auf (d.h. alle Merkmale müssen vorhanden sein damit intuitive Benutzung vorliegt) bzw. das Ausmaß intuitiver Benutzung spiegelt sich in allen Merkmalen gleichermaßen wider. Die in der Definition genannten Ausprägungen der Merkmale beschreiben somit eine vollkommen intuitive Benutzung (d.h. positives Extremum der jeweiligen Variable auf dem Kontinuum).

Eine Reihe von wissenschaftlichen Arbeiten (z.B. Blackler, 2018; Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2010; Ullrich, 2014) belegen die Übereinstimmung der IUUI-Forschergruppe und der QUT-Forschergruppe in dem Punkt, dass die kognitive Informationsverarbeitung während der Benutzung überwiegend unbewusst abläuft. Die QUT-Forschergruppe stützt dabei ihre Argumentation hauptsächlich auf das Drei-Ebenen-Modell von Rasmussen (1983) als theoretische Grundlage, was jedoch zu ähnlichen Einschränkungen wie bei der Verwendung der Handlungsregulationstheorie oder anderen Modellen der Informationsverarbeitung (z.B. D. A. Norman & Shallice, 1986) führt, so wie sie bereits im vorherigen Teilabschnitt 2.1.1 ausführlich erläutert wurden. Dementsprechend soll im Rahmen dieser Arbeit auch hier der Begriff „unbewusst“ als

überwiegend unbewusste kognitive Informationsverarbeitung auf Basis handlungsrelevanten Vorwissens an der Grenze zum menschlichen Bewusstsein (d.h. Fringe Consciousness) verstanden und damit verbundene Implikationen (siehe Mangan, 2015; E. Norman, 2017; E. Norman et al., 2010; Price & Norman, 2008; R. Reber et al., 2004; Schwarz, 2015; Zander et al., 2016) berücksichtigt werden. Auf diese Implikationen wird Rahmen einer genaueren Definition präattentiver metakognitiver Prozesse in Kapitel 2 eingegangen, da präattentive metakognitive Aspekte von der QUT in ihren Arbeiten nicht thematisiert werden. An dieser Stelle sollen stattdessen die theoretischen Grundlagen der QUT-Forschergruppe präsentiert, durch ausgewählte weiterführende Forschungsliteratur ergänzt, sowie mit der Handlungsregulationstheorie von Hacker (1986) verknüpft werden.

2.1.2.1 Zeitlich effiziente kognitive Informationsverarbeitung

Als zweites Merkmal intuitiver Benutzung beschreibt die QUT-Forschergruppe in ihren Arbeiten (z.B. Blackler, 2006, 2018; Blackler & Popovic, 2015; Blackler et al., 2010) die hohe zeitliche Effizienz intuitiver Benutzung, welche durch die hohe Geschwindigkeit bei der kognitiven Informationsverarbeitung aufgrund des überwiegend unbewussten Rückgriffs auf Vorwissen als Konsequenz intuitiver Benutzung entsteht (Baars, 1993; Bastick, 1982, 2003; Evans & Stanovich, 2013; Hammond, 1993; Hogarth, 2001; Price & Norman, 2008; Zajonc, 1980). Sie ist dadurch bewussten Prozessen in diesem Punkt überlegen. Blackler (2006) stützt dabei ihre Argumentation zum großen Teil auf Greenfield (2000), welcher beschreibt, dass das Gehirn eine halbe Sekunde braucht, bis ihm ein unbekannter Außenreiz bewusst ist und durch die Beanspruchung des Arbeitsgedächtnisses akkommodiert werden kann. Auf der anderen Seite benötigt das Gehirn bei unbewusster Verarbeitung und bei der Assimilation lediglich ein Hundertstel der Zeit (Greenfield, 2000), da Außenreize über offene Parameter mit hoher mentaler Effizienz holistisch unmittelbar in Schemas integriert werden können (Seel, 1991). Resultierendes sichtbares Verhalten wirkt daher im Vergleich zu überwiegend bewusster kognitiver Informationsverarbeitung eleganter, schneller, akkurater und sparsamer (Agor, 1986; Blackler, 2006, 2018; Blackler & Popovic, 2015; Greenfield, 2000), da für die kognitive Informationsverarbeitung vergleichsweise wenig Zeit aufgewendet werden muss und damit nur noch die physischen Operationen zeitlich ins Gewicht fallen.

Wie bereits zu Beginn dieses Kapitels in Teilabschnitt 2.1.1 angesprochen, stellt die zeitliche Effizienz bei der kognitiven Informationsverarbeitung ein weiteres Merkmal intuitiver Benutzung dar, da sie aus der hohen mentalen Effizienz der kognitiven Informationsverarbeitung resultiert. Sie sollte jedoch nicht nicht mit physischer Effizienz oder zeitlicher Effizienz der eigentlichen motorischen Handlungsausführung verwechselt werden (Blackler, 2006). Die zeitlich effiziente kognitive Verarbeitung kann sowohl objektiv als Leistungskriterium, als auch subjektiv in Form eines zufriedenstellenden mühelosen Gefühls (d.h. wahrgenommene geringe zeitliche Beanspruchung bei der kognitiven Informationsverarbeitung) erfasst werden (Blackler, 2018; Blackler et al., 2018; McAran, 2018; O'Brien, 2018).

Die QUT-Forschergruppe trifft darüber hinaus jedoch keine Aussagen zum Zusammenhang mit den anderen von der IUUI-Forschergruppe im Rahmen einer Handlungsperspektive vorgeschlagenen Usability-relevanten Leistungskriterien bzw. Konsequenzen intuiti-

ver Benutzung wie Effektivität, mentaler Effizienz und Zufriedenheit (siehe Teilabschnitt 2.1.1), obwohl beispielsweise der Zusammenhang zwischen zeitlicher und mental effizienter kognitiver Informationsverarbeitung, auf Basis der von der QUT-Forschergruppe verwendeten Literatur, naheliegend ist. Dementsprechend wird intuitive Benutzung von der QUT-Forschergruppe im Gegensatz zur IUUI-Forschergruppe nicht ganzheitlich aus einer Handlungsperspektive betrachtet, und sich von der Forschergruppe diesbezüglich lediglich auf den Aspekt der zeitlichen Effizienz bei der kognitiven Informationsverarbeitung als Konsequenz intuitiver Benutzung fokussiert - eine Vereinfachung, die jedoch im Widerspruch zur stets zielgerichteten Nutzung eines Systems und den damit verbundenen Leistungskriterien steht (siehe Teilabschnitt 2.1.1).

2.1.2.2 Geringe Verbalisierbarkeit der kognitiven Informationsverarbeitung

Als drittes Merkmal führt die QUT-Forschergruppe in diversen Arbeiten (z.B. Blackler, 2006, 2018; Blackler & Popovic, 2015; Blackler et al., 2010) die geringe explizite Verbalisierungsfähigkeit des Nutzers an, und stellt dieses Merkmal als wesentliches Zufriedenstellungskriterium für die subjektive Beurteilung intuitiver Benutzung dar. Laut der QUT-Forschergruppe hat intuitive Benutzung nur dann stattgefunden, wenn Menschen ihre eben durchgeführte Handlung aufgrund ihrer überwiegend unbewussten Anwendung von Vorwissen und der damit verbundenen unbewussten Assimilationsprozesse im Nachgang kaum beschreiben können (Blackler, 2006, 2018; Blackler & Popovic, 2015; M. L. Still & Still, 2018). Durch die daraus resultierende geringe Einsicht in den Verarbeitungsprozess, lässt sich an diesen Prozess auch kaum erinnern, er ist kaum nachvollziehbar oder explizit formal verbalisierbar (Agor, 1986; Bastick, 1982, 2003; Fischbein, 1987; Gigerenzer, 2007; Hammond, 1993; Klein, 2008; Price & Norman, 2008; Zander et al., 2016). Dies lässt sich von außen beispielsweise mithilfe eines parallel zur Handlung ausgeführten Think-Aloud-Protokolls objektiv beurteilen (siehe Blackler, 2018; Blackler et al., 2018).

Wenn intuitive Benutzung als überwiegend unbewusster Prozess auf Basis handlungsrelevanten Vorwissens an der Grenze zum menschlichen Bewusstsein stattfindet (Mangan, 2015; E. Norman, 2017; E. Norman et al., 2010; Price & Norman, 2008; Zander et al., 2016), stellt sich beim Nutzer in der Handlungsregulationsphase der Feedbackverarbeitung bei der bewussten Beurteilung des Handlungsfortschritts bzw. Handlungsergebnisses außerdem das subjektive Gefühl des „knowing without knowing why“ bzw. „knowing without being able to explain how we know“ (d.h. Feeling of Knowing; Gefühl der Kenntnis) ein (siehe Ackerman & Thompson, 2017; Koriat, 1993; Metcalfe, 2000; O'Brien et al., 2008; Price & Norman, 2008; Singer & Tiede, 2008; Topolinski, 2011; Ullrich, 2014). Dies ist der Fall da die kognitive Informationsverarbeitung überwiegend unbewusst war und der Nutzer diese somit nicht vollständig explizit verbalisieren (d.h. anhand verschiedener logischer Schritte explizit formulieren) kann, so wie es bei einer überwiegend bewussten kognitiven Informationsverarbeitung möglich wäre (Bastick, 2003; Gigerenzer, 2007; Hammond, 1993; Klein, 2008; Ullrich, 2014; Zander et al., 2016). Dieses Phänomen lässt sich mit der geringen expliziten Verbalisierungsfähigkeit von Gefühlen vergleichen, die Menschen in Bezug auf Objekte (z.B. Bilder) besitzen, wo alleine die explizite Formulierung Probleme bereitet (Nisbett & Wilson, 1977; Ullrich, 2014).

Dementsprechend kann die geringe explizite Verbalisierbarkeit der kognitiven Informationsverarbeitung nicht nur objektiv, sondern auch subjektiv in Form eines zufriedenstellenden Gefühls erfasst werden. Das Problem bei Verwendung von Verbalisierungsfähigkeit als charakterisierendes Merkmal intuitiver Benutzung ist jedoch, dass diese nicht nur sehr zwischen Personen (z.B. Experte vs. Nicht-Experte) variieren kann (Blackler et al., 2011), sondern Menschen auch bei vollständigem Nichtwissen sowie uninformiertem Trial-and-Error-Verhalten ihren kognitiven Denkprozess oftmals nicht verbalisieren können. Wie bereits im letzten Teilabschnitt angesprochen, sollte zufälliges Trial-and-Error-Verhalten, was zwar zu einem letztendlich erfolgreichen mental effizienten und auch für den Nutzer zufriedenstellenden Ergebnis führen kann, streng genommen nicht als intuitiv bezeichnet werden, da der Nutzer nicht zielgerichtet sein Vorwissen angewendet hat und er im Laufe des Prozesses bis zur Zielerreichung Fehler gemacht haben kann (siehe Ullrich, 2014). Dieser Umstand sollte daher bei Verwendung der Verbalisierbarkeit als Beurteilungskriterium intuitiver Benutzung immer berücksichtigt werden.

2.1.2.3 Merkmale aus Sicht der QUT-Forschergruppe

Zusammenfassend lassen sich unter Berücksichtigung der in diesem Abschnitt vorgestellten Literatur folgende charakterisierende Merkmale aus der Definition intuitiver Benutzung der QUT-Forschergruppe ableiten. Die überwiegend unbewusste kognitive Informationsverarbeitung auf Basis des mentalen Modells (d.h. handlungsrelevantem Vorwissen) bildet dabei die Vorbedingung ($\bullet$) und die anderen Merkmale die objektiven ($\rightarrow$) und die subjektiven ($\dashrightarrow$) Konsequenzen intuitiver Benutzung. Laut der QUT-Forschergruppe treten diese Merkmale bei einer intuitiven Benutzung gleichzeitig im gleichen Ausmaß auf bzw. das Ausmaß intuitiver Benutzung spiegelt sich in diesen Merkmalen gleichermaßen wider (Blackler, 2006). Die Merkmale wurden so formuliert, sodass sie die vollkommene intuitive Benutzung und somit das positive Extremum des Kontinuums repräsentieren. Je weiter man sich von den aufgelisteten Ausprägungen der Merkmale entfernt, umso geringer wird die intuitive Benutzung des Systems (z.B. weniger explizit verbalisierbar, weniger effektiv und weniger zufriedenstellend).

Charakteristische Merkmale intuitiver Benutzung aus Perspektive der QUT-Forschergruppe:

- $\bullet$ Überwiegend unbewusster kognitiver Informationsverarbeitungsprozess auf Basis von handlungsrelevantem Vorwissen an der Grenze zum menschlichen Bewusstsein

$\rightarrow$ Zeitlich effiziente kognitive Informationsverarbeitung

$\rightarrow$ Geringe explizite Verbalisierbarkeit der kognitiven Informationsverarbeitung

$\dashrightarrow$ Zufriedenstellende Benutzung

- Wahrgenommene geringe zeitliche Beanspruchung
- Wahrgenommene geringe explizite Verbalisierbarkeit

2.1.3 INTUI-Forschergruppe

Neben den beiden bereits vorgestellten HCI-Definitionen, welche intuitive Benutzung aus einer überwiegend objektiven Perspektive (d.h. objektive Merkmale: überwiegend unbe-

wusste Anwendung von handlungsrelevantem Vorwissen und damit verbundene Leistungsindikatoren wie Effektivität, mentale und zeitliche Effizienz) betrachten, konzentriert sich die INTUI-Forschergruppe (z.B. Diefenbach & Ullrich, 2015; Tretter, Diefenbach, & Ullrich, 2018; Ullrich, 2014; Ullrich & Diefenbach, 2010a, 2010b) in ihren Forschungsarbeiten überwiegend auf den Aspekt der Zufriedenstellung intuitiver Benutzung. Sie betont auf diese Weise besonders die mit intuitiver Benutzung verbundenen affektiven (d.h. gefühlsmäßigen, subjektiven) zufriedenstellenden Konsequenzen.

Laut Blackler und Popovic (2015) hat die Forschergemeinde dieser Perspektive zuvor wenig Bedeutung geschenkt, da aufgrund der überwiegend unbewussten Natur intuitiver Benutzung die allgemeine Vertrauenswürdigkeit von subjektiven Meinungen von Nutzern und damit der Aspekt der Zufriedenheit für dessen Beurteilung als relativ unzuverlässig eingestuft wurde. Wo die anderen beiden Forschergruppen als charakterisierende Merkmale auch objektive Charakteristika berücksichtigen und damit eine gute Balance zwischen objektiven und subjektiven Merkmalen anstreben, versucht die INTUI-Forschergruppe (z.B. Diefenbach & Ullrich, 2015; Tretter et al., 2018; Ullrich, 2014; Ullrich & Diefenbach, 2010a, 2010b) bei ihrer Auffassung intuitiver Benutzung überwiegend den Aspekt der User Experience (Hassenzahl & Tractinsky, 2006) in Form von subjektiven Empfindungen zu berücksichtigen (siehe Ullrich, 2014). Im Gegensatz zu den anderen beiden Forschergruppen stellt die INTUI-Forschergruppe keine Definition intuitiver Benutzung bereit, sondern beschreibt intuitive Benutzung in ihren Arbeiten anhand von verschiedenen Komponenten, also typische, im Zuge einer Handlung subjektiv wahrnehmbare Konsequenzen intuitiver Entscheidungen (Ullrich, 2014), die bereits in anderen Forschungsgebieten mehrfach identifiziert und empirisch belegt wurden (z.B. Betsch et al., 2011; Betsch & Glöckner, 2010; Gigerenzer, 2007; Glöckner & Witteman, 2010; Hogarth, 2001; Pfister, Jungermann, & Fischer, 2016; Plessner, Betsch, & Betsch, 2011):

1. Intuitives Handeln wird als schnell, mühelos und unmittelbar wahrgenommen (Beispiel: Ich „sehe“, dass Alternative A besser als Alternative B ist).
2. Der intuitive Prozess ist intransparent, das Ergebnis wird aber bewusst wahrgenommen (Beispiel: Ich weiß zwar nicht, „warum“ ich Alternative A präferiere, aber ich weiß, dass ich es tue).
3. Das Ergebnis intuitiven Handelns weist eine hohe wahrgenommene Validität auf (Beispiel: Ich bin mir „sehr sicher“, dass meine Intuition korrekt ist).

Diese subjektiv wahrnehmbaren Konsequenzen intuitiver Benutzung spiegeln sich laut der INTUI-Forschergruppe in den vier Komponenten Verbalisierungsfähigkeit, Mühelosigkeit, gefühlsgeleitetes Entscheiden (d.h. Bauchgefühl) und magischem Erleben wider. Diese treten im Gegensatz zur Auffassung der anderen beiden Forschergruppen nicht notwendigerweise alle gleichzeitig im gleichen Ausmaß bei einer intuitiven Benutzung auf (d.h. nicht alle Merkmale müssen vorhanden sein damit eine intuitive Benutzung vorliegt) bzw. das Ausmaß intuitiver Benutzung spiegelt sich nicht in allen Merkmalen gleichermaßen wider (siehe Ullrich, 2014). Stattdessen variieren diese Variablen in relativer Ausprägung zueinander auf einem Kontinuum (z.B. Verbalisierungsfähigkeit zeigt sich in einem Nutzungskontext stärker, wobei sich Mühelosigkeit kaum zeigt) und das Ausmaß an intuitiver Benutzung spiegelt sich in diesem Muster wider (Diefenbach & Ullrich, 2015; Tretter et al., 2018; Ullrich, 2014).

Wie bereits die QUT-Forschergruppe zuvor, nimmt die INTUI-Forschergruppe auch keine explizite Aufteilung in Vorbedingung und Konsequenzen intuitiver Benutzung vor und stellt darüber hinaus keine explizite Definition intuitiver Benutzung bereit. Es lassen sich jedoch alle vier Komponenten als Konsequenzen von intuitiver Benutzung interpretieren. Darüber hinaus wird ein überwiegend unbewusster kognitiver Informationsverarbeitungsprozess auf Basis handlungsrelevanten Vorwissens in der Forschungsliteratur der INTUI-Forschergruppe in Zusammenhang mit diesen Komponenten als Voraussetzung intuitiver Benutzung diskutiert (siehe Diefenbach & Ullrich, 2015; Ullrich, 2013, 2014; Ullrich & Diefenbach, 2010a). Dieser Prozess kann daher als Vorbedingung intuitiver Benutzung betrachtet werden. Die vier Komponenten werden nun nachfolgend vorgestellt und anhand weiterer Arbeiten anderer HCI-Forscher (z.B. Antle, Corness, & Droumeva, 2009; Macaranas, 2013; Macaranas, Antle, & Riecke, 2015; McAran, 2018; O'Brien et al., 2008; M. L. Still & Still, 2018), sowie damit verbundener Grundlagenliteratur aus anderen Forschungsgebieten ergänzt und mit der Handlungsregulationstheorie von Hacker (1986) verknüpft. Dies ist sinnvoll, da die Ausführungen der INTUI-Forschergruppe bezüglich des genauen Zusammenspiels der einzelnen Komponenten nicht sehr ins Detail gehen und man so die Komponenten im Rahmen dieser Arbeit stärker in Beziehung zu den bereits genannten Merkmalen der anderen beiden Forschergruppen setzen möchte.

2.1.3.1 Verbalsierungsfähigkeit und Mühelosigkeit

Eine geringe explizite Verbalisierungsfähigkeit stellt auch bei der QUT-Forschergruppe ein charakterisierendes Merkmal intuitiver Benutzung dar und wurde von der INTUI-Forschergruppe als Komponente aus der Definition der QUT-Forschergruppe dementsprechend übernommen (Diefenbach & Ullrich, 2015; Ullrich, 2014; Ullrich & Diefenbach, 2010a, 2010b). Sie kann sowohl als objektives Leistungskriterium als auch als subjektive Konsequenz intuitiver Benutzung aufgefasst werden. Die Komponente Mühelosigkeit beschreibt die allgemein hohe Mühelosigkeit, die man bei einer intuitiven Benutzung subjektiv wahrnimmt (Diefenbach & Ullrich, 2015; Ullrich, 2014; Ullrich & Diefenbach, 2010a, 2010b). Diese Mühelosigkeit sollte sich demzufolge objektiv in der effektiven, mental und zeitlich effizienten kognitiven Informationsverarbeitung widerspiegeln, die im Zuge einer überwiegend unbewussten Anwendung von handlungsrelevantem Vorwissen aus den damit verbundenen kognitiven Assimilationsprozessen der Handlungsregulation resultiert. Aus einer subjektiven Erlebnissicht manifestiert sich intuitive Benutzung daher in Form spontaner und unmittelbarer Ideen und Einfälle, die nicht vorsätzlich hervorgerufen oder ignoriert werden können (d.h. Assimilation läuft unbewusst ohne willentliche Steuerung ab). Intuitive Benutzung wird infolgedessen auch subjektiv als generell mühelos erlebt (siehe McAran, 2018; Price & Norman, 2008; R. Reber & Schwarz, 2001; Topolinski & Strack, 2009; Ullrich, 2014). Die Komponente der Mühelosigkeit bildet auf diese Weise einen subjektiven Zufriedenheitsaspekt intuitiver Benutzung ab, und muss deswegen auch als Zufriedenheitsmerkmal berücksichtigt werden (Blackler et al., 2018; McAran, 2018).

Die INTUI-Forschergruppe bezeichnet dieses Merkmal aufgrund des fehlenden expliziten Bezugs zu Usability zwar nicht wie die IUUI-Forschergruppe als Zufriedenstellung, stimmt jedoch hauptsächlich mit der Auffassung der IUUI-Forschergruppe überein, welche ja auch

die Konsequenzen intuitiver Benutzung (z.B. mentale Effizienz) als Zufriedenheitsmerkmale (z.B. wahrgenommene geringe mentale Beanspruchung) berücksichtigen. Lediglich der bei der Definition der IUUI-Forschergruppe fehlende Aspekt der zeitlichen Effizienz bei der kognitiven Informationsverarbeitung wurde von der INTUI-Forschergruppe in der Beschreibung ihrer Komponente Mühelosigkeit noch zusätzlich als Gefühl berücksichtigt und von der QUT-Forschergruppe übernommen (Diefenbach & Ullrich, 2015; Tretter et al., 2018; Ullrich, 2014; Ullrich & Diefenbach, 2010a, 2010b).

2.1.3.2 Gefühlsgeleitetes Entscheiden und magisches Erleben

Im Gegensatz zur hohen Mühelosigkeit und geringen Verbalisierungsfähigkeit, wo sich beide Komponenten in objektiver und subjektiver Form als Konsequenzen intuitiver Benutzung interpretieren und festhalten lassen, versucht die INTUI-Forschergruppe mit der Komponente Bauchgefühl zusätzlich ausschließlich subjektiv zu erfassen, wie sich die kognitive Informationsverarbeitung bei der Anwendung des handlungsrelevanten Vorwissens ganzheitlich „gefühlsleitend“ anfühlt und die intuitive Benutzung dabei von Gefühlen „gesteuert“ wird (Diefenbach & Ullrich, 2015; Tretter et al., 2018; Ullrich, 2014; Ullrich & Diefenbach, 2010a, 2010b). In Anlehnung an Hammond (1993) fasst die INTUI-Forschergruppe intuitive Benutzung auch als einen kognitiven Informationsverarbeitungsprozess auf, der auf irgendeine Weise eine Handlungstendenz (d.h. Lösung, Antwort, Bauchgefühl) liefert, ohne dass einem dabei die konkrete Lösungsfindung bewusst ist. Für den Handelnden resultiert daraus, dass auch das Ergebnis des damit verbundenen kognitiven Prozesses, die durch Intuition erlangte Erkenntnis und das damit verbundene Gefühl, schwer nachzuvollziehen oder logisch zu verteidigen ist (Diefenbach & Ullrich, 2015; Ullrich, 2014). Die erlangte Erkenntnis bzw. Selbstsicherheit zeichnet sich durch eine hohe subjektive Gewissheit aus, denn auch wenn sich der Handelnde nicht erklären kann, woher seine Einsicht kommt, er ist sich trotzdem sicher, auf deren Basis die richtige Entscheidung zu treffen (Ullrich, 2014).

Im Zuge der kognitiven Informationsverarbeitung kommt es häufig zu Situationen, in denen die unbewusste Assimilation neuer Informationen durch existierende Schemas aufgrund der Neuartigkeit der Situation (d.h. schemainkonsistente Situation) nicht mehr möglich ist und der Handelnde zur bewussten Verarbeitung (d.h. Akkommodation) wechseln sollte, um seine effektive Zielerreichung nicht zu gefährden (siehe Betsch et al., 2011; Kopp & Mandl, 2005; O'Brien et al., 2008; Saifoulline & Hemberger, 2011; Seel, 1991). In den Arbeiten der INTUI-Forschergruppe (siehe Ullrich, 2014) und bei O'Brien et al. (2008) wird für den Vorgang der Assimilation der Begriff *Feedforward* in Anlehnung an Basso und Belardinelli (2006) verwendet. Feedforward ist ein neurologischer Mechanismus, welcher basierend auf dem erwarteten zukünftigen Soll-Zustand in einem bestimmten Handlungskontext bis zum einem Zielzustand mit hoher mentaler Effizienz unbewusst fortschreitet, bis es zur einer Situation kommt, welche ein bewusstes Einschreiten und die Auswahl des nächsten Handlungsschritts erfordert (O'Brien et al., 2008).

Laut einer Reihe von wissenschaftlichen Arbeiten (z.B. Ackerman & Thompson, 2017; Epstein, 2010; Hodgkinson, Langan-Fox, & Sadler-Smith, 2008; Mangan, 2015; E. Norman, 2017; Price & Norman, 2008; R. Reber & Schwarz, 2001; R. Reber et al., 2004; Schwarz, 2015; Stanovich et al., 2014; V. Thompson & Morsanyi, 2012; V. Thompson et al., 2011;

Topolinski & Strack, 2009) kommt es in solchen Situationen, die sich auf der Grenze des menschlichen Bewusstseins (d.h. Fringe Consciousness) abspielen, zu einem kurzen unbewussten affektiven Richtimpuls (d.h. Fringe), dessen Stärke von der Flüssigkeit des Assimilationsprozesses (d.h. Fluency) bestimmt ist. Im Forschungsbereich intuitiver Benutzung greifen im Gegensatz zur INTUI-Forschergruppe M. L. Still und Still (2018) diesen präattentiven metakognitiven Aspekt zwar indirekt auf, sprechen aber nicht explizit davon als Flüssigkeit, obwohl sie damit verbundene Literatur (z.B. Betsch, 2008; Epstein, 2010; Hodgkinson et al., 2008; Topolinski & Strack, 2009) referenzieren.

Flüssigkeit ist zum einen die subjektive Erfahrung von Mühelosigkeit (d.h. subjektive Erfahrung objektiver Leistungskriterien: hohe wahrgenommene Effektivität, hohe wahrgenommene zeitliche und mentale Effizienz bei der kognitiven Informationsverarbeitung), mit der eine Person kognitiv Informationen verarbeiten kann (siehe R. Reber et al., 2004), die mithilfe zufriedenstellender (d.h. subjektiver) Konsequenzen intuitiver Benutzung bereits bei den anderen beiden Forschergruppen indirekt berücksichtigt ist. Zum anderen versteht man darunter auch aus einer objektiven Perspektive eine hohe mentale und zeitliche Effizienz bei der effektiven kognitiven Informationsverarbeitung (d.h. als objektive Leistungskriterien bzw. Konsequenzen intuitiver Benutzung) (siehe R. Reber et al., 2004), die durch Leistungskriterien bzw. objektive Konsequenzen intuitiver Benutzung bereits von den anderen beiden Forschergruppen indirekt berücksichtigt ist.

R. Reber et al. (2004) bezeichnen den ersten Aspekt als die subjektive Dimension von Flüssigkeit (d.h. Gefühl von Flüssigkeit oder subjektive Flüssigkeit) und den zweiten Aspekt als deren objektive Dimension (d.h. objektive Flüssigkeit). Die subjektive Flüssigkeit wird oftmals in der Forschungsliteratur auch als Gefühl von Richtigkeit bezeichnet (z.B. Ackerman & Thompson, 2017; V. Thompson, 2009; V. Thompson et al., 2011). Im Rahmen dieser Arbeit wird jedoch einheitlich vom Gefühl von Flüssigkeit oder subjektiver Flüssigkeit gesprochen, da mit beiden Bezeichnungen ein Gefühl von Einfachheit gemeint ist (Ackerman & Thompson, 2017; R. Reber, Fazendeiro, & Winkielman, 2002; R. Reber et al., 2004; V. Thompson, Turner, Pennycook et al., 2013), das die Qualität der Informationsverarbeitung widerspiegelt (d.h. es fühlt sich „richtig“ oder „einfach“ an, wenn die Informationsverarbeitung effektiv, mental und zeitlich effizient war; es fühlt sich „falsch“ oder „schwer“ an, wenn die Informationsverarbeitung nicht effektiv, nicht mental und nicht zeitlich effizient war).

Je flüssiger der Assimilationsprozess oder das Feedforward abläuft (abhängig von z.B. dem Grad der Richtigkeit und Differenziertheit des handlungsrelevanten Vorwissens, den allgemeinen Gesetzmäßigkeiten des Nutzungskontextes), umso sicherer ist sich der Handelnde die neuartige Situation bewältigen zu können (siehe Ackerman & Thompson, 2017; E. Norman, 2017; O'Brien et al., 2008; Schwarz, 2015; J. D. Still, Still, & Grgic, 2015; M. L. Still & Still, 2018; V. Thompson, 2009; V. Thompson et al., 2011). Ein positiver affektiver Richtimpuls signalisiert dem Handelnden, dass die Assimilation und das damit verbundene mentale Modell zur Bewältigung der neuartigen Situation ausreichend ist und keine Notwendigkeit zur Anpassung des mentalen Modells durch Akkommodation besteht (oder nur minimale Anpassungen, die keinen Wechsel nötig machen), was eine bewusste kognitive Informationsverarbeitung nötig machen würde. Es entsteht dabei ein *globales, mikrobewusstes, sicheres, zufriedenes Bauchgefühl* in Form eines Gefühls von

Flüssigkeit (d.h. subjektive Flüssigkeit), welches durch Mikrobewusstsein auf verschiedenen Stufen der neuronalen Verarbeitung einen kondensierten Informationsüberblick über nicht bewusst zugängliche Informationen und eine starke Handlungstendenz liefert, ohne bewussten Zugriff auf die Vorläufer dieses Gefühls zu bieten (siehe Ackerman & Thompson, 2017; Gigerenzer, 2007; Hammond, 1993; Mangan, 2015; Price & Norman, 2008; R. Reber et al., 2002; V. Thompson, 2009; V. Thompson & Morsanyi, 2012; V. Thompson et al., 2011; Topolinski & Strack, 2009).

Aufgrund dieses subjektiven Gefühls von Flüssigkeit oder Richtigkeit, welches als Basis von präattentiven metakognitiven Prozessen (siehe Teilabschnitt 2.1.1.2) über den Wechsel zwischen unbewusster und bewusster kognitiver Verarbeitung „gefühlsmäßig“ entscheidet (siehe Ackerman & Thompson, 2017; Stanovich et al., 2014), also dementsprechend die kognitive Informationsverarbeitung für den Handelnden intransparent leitet, wird intuitive Benutzung auch als eher vage, nicht nachvollziehbar, unkoordiniert und gefühlsgeleitet im Sinne eines Bauchgefühls erlebt (Price & Norman, 2008). Da der Handelnde sich die Quelle seiner erhöhten wahrgenommen Validität nur schwer erklären und verbalisieren kann (Hammond, 1993), kann sich dieses positive Gefühl für den Handelnden nahezu „magisch“ anfühlen. Dieses magische Erleben wird von der INTUI-Forschergruppe als weiteres subjektives Merkmal intuitiver Benutzung anerkannt und kann somit auch als subjektive Konsequenz intuitiver Benutzung angesehen werden (siehe Diefenbach & Ullrich, 2015; Ullrich, 2013, 2014). Diese Intransparenz des kognitiven Informationsverarbeitungsprozesses bei der Vorwissensaktivierung und die unbekannte Herkunft des Vorwissens (Klein, 1998) kann laut der INTUI-Forschergruppe (siehe Ullrich, 2014) und anderen Forschern (siehe Blackler, 2018; Stillman, Shen, & Ferguson, 2018) nicht nur zur mangelnden expliziten Verbalisierbarkeit des Prozesses, sondern auch zu einer zusätzlichen Mystifizierung von intuitiver Benutzung führen (Ullrich, 2014). Intuition kann deswegen als übernatürliches Geschenk oder als Eingebung erscheinen (Zander et al., 2016). Der subjektive Zufriedenheitsaspekt intuitiver Benutzung lässt sich daher nicht nur am Bauchgefühl, der geringen Verbalisierungsfähigkeit oder der hohen Mühelosigkeit subjektiv bewerten, sondern auch am magischen Erleben der intuitiven Benutzung selbst.

Im Gegensatz zu den anderen beiden Forschergruppen betont die INTUI-Forschergruppe bei der Vorstellung ihrer Komponenten explizit, dass die einzelnen charakterisierenden Merkmale intuitiver Benutzung nicht nur im Sinne eines Kontinuums gleichzeitig in ihrer konkreten Ausprägung variieren können, sondern dass diese sogar nicht immer alle gleichzeitig vorhanden sein müssen, damit ein Produkt, Dienst oder Technologie als intuitiv wahrgenommen wird. Es ergeben sich für verschiedene Produktkategorien verschiedene Muster (d.h. intuitive Benutzung äußert sich bei manchen Produkten bezüglich Verbalisierungsfähigkeit stark und bei manchen nicht oder weniger). Ullrich (2014) liefert hierzu beispielsweise empirische Befunde, die bezüglich der vorgestellten Komponenten deutliche Unterschiede zwischen Unterhaltungselektronik und Haushaltsgeräten zeigen. Die Interpretation der relativen Ausprägung der einzelnen Merkmale zueinander werden von der INTUI-Forschergruppe als INTUI-Patterns bezeichnet. Mit diesen lässt sich das Ausmaß der persönlich erlebten intuitiven Benutzung mit einem Produkt aus einer Erlebnissicht beschreiben (Diefenbach & Ullrich, 2015; Ullrich, 2014; Ullrich & Diefenbach, 2010b). Das Ausmaß intuitiver Benutzung muss sich demnach nicht in allen Komponenten gleichzeitig und im gleichen Ausmaß widerspiegeln, damit eine intuitive Benutzung vorliegt.

Es soll an dieser Stelle entsprechend festgehalten werden, dass mit allen von der INTUI-Forschergruppe vorgestellten subjektiven Komponenten Mühelosigkeit, Verbalisierungsfähigkeit, Bauchgefühl und magischem Erleben immer bewusst wahrnehmbare Gefühle und damit subjektive Konsequenzen intuitiver Benutzung adressiert werden, die unter Berücksichtigung des Flüssigkeitskonzepts und speziell der Arbeit von R. Reber et al. (2002) als phänomenologische Erfahrungen bzw. Konsequenzen der objektiven Flüssigkeit während des Assimilationsprozesses und in Abhängigkeit des Nutzungskontextes auftreten. Jede Komponente konzentriert sich dabei auf eine andere subjektive Konsequenz intuitiver Benutzung und versucht dabei die subjektiven Auswirkungen einer hohen objektiven Flüssigkeit während des Assimilationsprozesses von einer anderen Seite aus zu beleuchten (siehe Ackerman & Thompson, 2017; R. Reber et al., 2002; Topolinski & Strack, 2009). Hinter dem aus diesen verschiedenen Erfahrungen resultierenden umfassenden Gefühl von Flüssigkeit steckt jedoch keine „Magie“. Viel mehr unterstützt oder hemmt dieses Gefühl von Flüssigkeit abhängig vom Handlungskontext als präattentives metakognitives Gefühl den Zugang zu handlungsrelevantem Vorwissen (Sinclair & Ashkanasy, 2005; Sinclair, Ashkanasy, Chattopadhyay, & Boyle, 2002). Dies kann zum einen hedonische Folgen haben, da ein müheloses Abrufen von Vorwissen stets als vertraut, angenehm, ungefährlich und emotional positiv erlebt wird (Winkielman & Cacioppo, 2001). Zum anderen kann es auch epistemische Folgen haben, denn es wird als valide und „wahr“ erlebt (Morewedge & Kahneman, 2010; Pfister et al., 2016; Zajonc & Markus, 1982; Zander et al., 2016).

Intuitive Benutzung besitzt daher nicht nur eine subjektive Gefühlskomponente, die als subjektiver Zufriedenheitsaspekt in einer Definition berücksichtigt werden sollte, sondern liefert mit dem Gefühl von Flüssigkeit zusätzlich auch ein metakognitives Gefühl (siehe Ackerman & Thompson, 2017; O’Brien et al., 2008; V. Thompson, 2009; V. Thompson et al., 2011; V. Thompson, Turner, Pennycook et al., 2013). Wie bereits im letzten Teilabschnitt 2.1.1 angesprochen, beschreibt der Begriff *Metakognition* im Allgemeinen Prozesse, durch die Menschen sich mit ihren eigenen kognitiven Informationsverarbeitungsprozessen befassen und durch bestimmte Aspekte dieser Prozesse in ihrem Denken und Verhalten beeinflusst werden (Betsch et al., 2011). Im Zuge der überwiegend unbewussten kognitiven Informationsverarbeitung auf Basis handlungsrelevanten Vorwissens an der Grenze zum menschlichen Bewusstsein (d.h. Fringe Consciousness) beeinflusst das Gefühl von Flüssigkeit als mikrobewusstes, präattentives metakognitives Gefühl das intuitive Handeln, da die Flüssigkeit einen affektiven Richtimpuls erzeugt und dadurch grundsätzlich bestimmt, wann der Handelnde zwischen unbewusster und bewusster Verarbeitung mithilfe seiner exekutiven Funktionen wechseln muss. Auf diese Weise können Menschen ihre Gedanken und ihr Verhalten basierend auf diesem Gefühl flexibel überwachen und regulieren (Price & Norman, 2008). Das Gefühl von Flüssigkeit reflektiert somit ganzheitlich die Tiefe der intuitiven Verarbeitung und die damit verbundene subjektive Zufriedenheit (Ackerman & Thompson, 2017; Price & Norman, 2008; R. Reber et al., 2002; R. Reber et al., 2004; V. Thompson et al., 2011; V. Thompson, Turner, Pennycook et al., 2013).

Eine Reihe von wissenschaftlichen Arbeiten konnten bereits demonstrieren, dass sich der affektive Richtimpuls bzw. die Flüssigkeit während der Assimilation in einer Reihe weiterer phänomenologischer Erfahrungen wie Vertrautheit, Kenntnis, Zielerreichung und Lernaufwand manifestiert (Ackerman & Thompson, 2017; Price & Norman, 2008; R. Reber et al., 2002; R. Reber et al., 2004), so wie sie beispielsweise teilweise von den drei Forschungsgruppen als subjektive Indikatoren einer zufriedenstellenden intuitiven Benutzung

beschrieben werden. Ein Überblick über weitere Gefühle ist ausführlich bei Ackerman und Thompson (2017) zu finden, wobei die genaue Phänomenologie dieser Gefühle noch weiter empirisch fundiert werden muss (Ackerman & Thompson, 2017; Price & Norman, 2008; V. Thompson, 2009; V. Thompson et al., 2011).

Nichtsdestotrotz lässt sich das allgemeine zufriedenstellende metakognitive Gefühl von Flüssigkeit als umfassendes Gefühl festhalten (siehe O'Brien et al., 2008), worunter sich entsprechend auch die von der IUUI-Forschergruppe (siehe Teilabschnitt 2.1.1) und der QUT-Forschergruppe (siehe Teilabschnitt 2.1.2) vorgeschlagenen subjektiven Indikatoren einer zufriedenstellenden intuitiven Benutzung gruppieren lassen. Auf den genauen Wirkmechanismus von Flüssigkeit und auf präattentive Metakognition wird erst im folgenden Abschnitt 2.2 genauer eingegangen, da im Forschungsfeld zu intuitiver Benutzung lediglich O'Brien et al. (2008) einen metakognitiven Aspekt im Bereich intuitiver Benutzung explizit ansprechen (d.h. beziehen sich explizit auf ein Gefühl der Kenntnis und Selbstsicherheit, die phänomenologisch in Abhängigkeit des Nutzungskontextes aus objektiver Flüssigkeit resultieren können, siehe Ackerman & Thompson, 2017). Darüber hinaus beschreibt die INTUI-Forschergruppe zwar die gefühlsgeleitete Natur von intuitiven Entscheidungen in ihren Arbeiten, konzentriert sich in ihren Ausführungen aber hauptsächlich auf die Beschreibung des Gefühlsaspekts intuitiver Benutzung und nicht explizit auf dessen präattentiven metakognitiven Wirkmechanismus (siehe Diefenbach & Ullrich, 2015; Ullrich, 2013, 2014).

2.1.3.3 Merkmale aus Sicht der INTUI-Forschergruppe

Zusammenfassend lassen sich unter Berücksichtigung der in diesem Abschnitt vorgestellten Literatur folgende charakterisierende Merkmale aus den genannten Komponenten intuitiver Benutzung der INTUI-Forschergruppe ableiten. Die überwiegend unbewusste kognitive Informationsverarbeitung auf Basis des mentalen Modells (d.h. handlungsrelevantem Vorwissen) lässt dabei als Vorbedingung ($\bullet$) und die anderen Merkmale als die objektiven ($\rightarrow$) und die subjektiven ($\dashrightarrow$) Konsequenzen intuitiver Benutzung interpretieren. Laut der INTUI-Forschergruppe treten diese Merkmale nicht notwendigerweise immer alle gleichzeitig auf (d.h. nicht alle Merkmale müssen vorhanden sein damit intuitive Benutzung vorliegt) bzw. das Ausmaß intuitiver Benutzung spiegelt sich in diesen Merkmalen nicht immer gleichermaßen wider. Stattdessen beschreibt die relative Ausprägung der Merkmale zueinander das Ausmaß an intuitiver Benutzung (Ullrich, 2014). Die Merkmale wurden so formuliert, sodass sie die vollkommene intuitive Benutzung und somit das positive Extremum des Kontinuums repräsentieren. Je weiter man sich von den aufgelisteten Ausprägungen der Merkmale entfernt, umso geringer wird die intuitive Benutzung des Systems (z.B. weniger wahrgenommenes magisches Erleben, weniger wahrgenommenes gefühlsgeleitetes Entscheiden), wobei hier die relative Ausprägung der Komponenten zusätzlich bei der Interpretation berücksichtigt werden muss.

Im Gegensatz zu den anderen beiden Forschergruppen stellt die INTUI-Forschergruppe keine explizite Definition intuitiver Benutzung bereit, sondern beschreibt intuitive Benutzung lediglich anhand von vier Komponenten. Da dies einen höheren Interpretationsspielraum zulässt und sich somit eine Einteilung in Vorbedingung, sowie subjektive und objektive

Konsequenzen intuitiver Benutzung nur indirekt mit Zuhilfenahme weiterer Literatur vornehmen lässt, wurde diese Interpretationsfreiheit bei der Ableitung der obigen Merkmale (d.h. direkte oder indirekt aus den Komponenten abgeleitet) kenntlich gemacht. Auf diese Weise konnten die von der INTUI-Forschergruppe vorgeschlagenen Merkmale mit den Merkmalen der anderen beiden Forschergruppen in Verbindung gesetzt werden.

Charakteristische Merkmale intuitiver Benutzung aus Perspektive der INTUI-Forschergruppe:

- Überwiegend unbewusster kognitiver Informationsverarbeitungsprozess auf Basis von handlungsrelevantem Vorwissen an der Grenze zum menschlichen Bewusstsein (indirekt aus allen vier Komponenten abgeleitet, da dies die Voraussetzung für alle Komponenten darstellt)

→ Zeitlich effiziente kognitive Informationsverarbeitung (indirekt aus der Komponente Mühelosigkeit abgeleitet)

→ Mental effiziente kognitive Informationsverarbeitung (indirekt aus der Komponente Mühelosigkeit abgeleitet)

→ Effektive Benutzung (indirekt aus der Komponente Mühelosigkeit abgeleitet)

→ Geringe explizite Verbalisierbarkeit der kognitiven Informationsverarbeitung (direkt aus der Komponente Verbalisierungsfähigkeit abgeleitet)

⇢ Zufriedenstellende Benutzung

- Wahrgenommene geringe mentale Beanspruchung (indirekt aus der Komponente Mühelosigkeit abgeleitet)

- Wahrgenommene geringe zeitliche Beanspruchung (indirekt aus der Komponente Mühelosigkeit abgeleitet)

- Wahrgenommene hohe Zielerreichung (indirekt aus der Komponente Mühelosigkeit abgeleitet)

- Wahrgenommene geringe Fehlerrate (indirekt aus der Komponente Mühelosigkeit abgeleitet)

- Wahrgenommene geringe explizite Verbalisierbarkeit (direkt aus der Komponente Verbalisierbarkeit abgeleitet)

- Wahrgenommenes gefühlsgeleitetes Entscheiden (direkt aus der Komponente Bauchgefühl abgeleitet)

- Wahrgenommenes magisches Erleben (direkt aus der Komponente magisches Erleben abgeleitet)

2.1.4 Merkmale aus Sicht aller HCI-Forschergruppen

Auf Basis der dargestellten theoretischen Überlegungen und empirischen Befunde assoziiert die Forschungsgemeinschaft in der HCI (d.h. vorgestellte Forschergruppen, sowie unabhängige Forscher) in ihren Veröffentlichungen die folgenden charakterisierenden Merkmale mit intuitiver Benutzung entweder direkt (d.h. werden in der Definition einer bestimmten Forschergruppe explizit genannt und können daher direkt als Merkmale abgeleitet werden) oder indirekt (d.h. werden in der Definition einer bestimmten Forschergruppe nicht explizit genannt und können daher nur aus den explizit genannten Merkmalen implizit abgeleitet

werden). Die überwiegend unbewusste kognitive Informationsverarbeitung auf Basis des mentalen Modells (d.h. handlungsrelevantem Vorwissen) kann dabei als Vorbedingung (•) und die anderen Merkmale als die objektiven (→) und die subjektiven (⇢) Konsequenzen intuitiver Benutzung interpretiert werden.

Charakteristische Merkmale intuitiver Benutzung:

- • Überwiegend unbewusster kognitiver Informationsverarbeitungsprozess auf Basis von handlungsrelevantem Vorwissen (alle Forschergruppen und unabhängige Forscher, z.B. Blackler, 2006, 2018; Blackler et al., 2018; Diefenbach & Ullrich, 2015; Hurtienne, 2011; Macaranas, 2013; Mohs, Hurtienne, Kindsmüller et al., 2006; Naumann et al., 2007; O'Brien et al., 2008; Ullrich, 2014)
- → Zeitlich effiziente kognitive Informationsverarbeitung (QUT-, INTUI-Forschergruppe und unabhängige Forscher, z.B. Blackler, 2006, 2018; Blackler et al., 2018; Blackler et al., 2010; Diefenbach & Ullrich, 2015; Reinhardt et al., 2018; Ullrich, 2014)
- → Effektive Benutzung (IUUI-, INTUI-Forschergruppe und unabhängige Forscher, z.B. Hurtienne, 2011; Mohs, Hurtienne, Kindsmüller et al., 2006; Naumann et al., 2007; O'Brien et al., 2008; O'Brien et al., 2010; Reinhardt et al., 2018; M. L. Still & Still, 2018; Ullrich, 2014)
- → Mental effiziente kognitive Informationsverarbeitung (IUUI-, INTUI-Forschergruppe und unabhängige Forscher, z.B. Hurtienne, 2011; Mohs, Hurtienne, Kindsmüller et al., 2006; Naumann et al., 2007; Reinhardt & Hurtienne, 2017; Ullrich, 2014)
- → Geringe Verbalisierbarkeit der kognitiven Informationsverarbeitung (QUT-, INTUI-Forschergruppe und unabhängige Forscher, z.B. Blackler, 2006; Blackler et al., 2010; M. L. Still & Still, 2018; Tretter et al., 2018; Ullrich, 2014)
- ⇢ Zufriedenstellende Benutzung (alle Forschergruppen, z.B. Blackler, 2006; Hurtienne, 2011; Ullrich, 2014) durch starkes metakognitives Gefühl von Flüssigkeit (z.B. O'Brien et al., 2008)

  - Wahrgenommene geringe mentale Beanspruchung (IUUI-, INTUI-Forschergruppe und unabhängige Forscher, z.B. Blackler et al., 2018; Hurtienne, 2011; Naumann & Hurtienne, 2010; O'Brien, 2018; Reinhardt et al., 2018; Ullrich, 2014)

  - Wahrgenommene hohe Zielerreichung (IUUI-, INTUI-Forschergruppe und unabhängige Forscher, z.B. Blackler et al., 2018; Hurtienne, 2011; Naumann & Hurtienne, 2010; Reinhardt et al., 2018; Ullrich, 2014)

  - Wahrgenommene geringe zeitliche Beanspruchung (QUT-, INTUI-Forschergruppe und unabhängige Forscher, z.B. Blackler, 2018; Diefenbach & Ullrich, 2015; McAran, 2018; Ullrich, 2014)

  - Wahrgenommene hohe Vertrautheit (IUUI-Forschergruppe und unabhängige Forscher, z.B. Blackler, 2018; Blackler et al., 2018; Hurtienne, 2011; Naumann & Hurtienne, 2010; Reinhardt et al., 2018)

  - Wahrgenommener geringer Lernaufwand (IUUI-Forschergruppe und unabhängige Forscher, z.B. Blackler, 2018; Blackler et al., 2018; Hurtienne, 2011; Naumann & Hurtienne, 2010; Reinhardt et al., 2018; M. L. Still & Still, 2018)

- Wahrgenommene geringe Fehlerrate (IUUI-, INTUI-Forschergruppe und unabhängige Forscher, z.B. Blackler, 2018; Blackler et al., 2018; Hurtienne, 2011; Naumann & Hurtienne, 2010; Reinhardt et al., 2018; Ullrich, 2014)
- Wahrgenommene geringe Verbalisierbarkeit (QUT-, INTUI-Forschergruppe und unabhängige Forscher, z.B. Blackler, 2006; Blackler et al., 2010; M. L. Still & Still, 2018; Tretter et al., 2018; Ullrich, 2014)
- Wahrgenommenes gefühlsgeleitetes Entscheiden (INTUI-Forschergruppe, z.B. Diefenbach & Ullrich, 2015; Tretter et al., 2018; Ullrich, 2013, 2014)
- Wahrgenommenes magisches Erleben (INTUI-Forschergruppe, z.B. Diefenbach & Ullrich, 2015; Tretter et al., 2018; Ullrich, 2013, 2014)

Laut der IUUI- und QUT-Forschergruppe variieren die von ihnen vorgeschlagenen Merkmale (alle obigen Merkmale außer wahrgenommenes gefühlsgeleitetes Entscheiden und wahrgenommenes magisches Erleben) als Variablen bei einer intuitiven Benutzung immer gleichzeitig im gleichen Ausmaß (d.h. alle von den beiden Forschergruppen vorgeschlagenen Merkmale müssen bei einer intuitiven Benutzung in gleichem Ausmaß vorhanden sein) bzw. das Ausmaß intuitiver Benutzung spiegelt sich in allen von den beiden Forschergruppen vorgeschlagenen Merkmalen gleichermaßen auf einem einem Kontinuum wider (Anmerkung: Einige Merkmale unterscheiden sich jedoch zwischen der IUUI- und QUT-Forschergruppe, siehe Teilabschnitt 2.1.1 und 2.1.2).

Im Gegensatz zu den anderen Forschergruppen ist die INTUI-Forschergruppe bezüglich der von ihr vorgeschlagenen vier Komponenten und damit auch den daraus überwiegend indirekt abgeleiteten Merkmalen anderer Ansicht. Die von der INTUI-Forschergruppe vorgeschlagenen Merkmale (siehe Teilabschnitt 2.1.3) treten nicht alle notwendigerweise gleichzeitig und in gleicher Ausprägung, sondern in relativer Abhängigkeit vom jeweiligen Nutzungskontext während intuitiver Benutzung auf (d.h. nicht alle aus den Komponenten der Forschergruppe abgeleiteten Merkmale müssen immer gleichzeitig auftreten damit intuitive Benutzung vorliegt). Das Ausmaß der Merkmale variiert dabei auf einem Kontinuum (d.h. intuitive Benutzung spiegelt sich in vorhandenen Merkmalen unterschiedlich wider und variiert damit gemäß seiner relativen Ausprägung auch im Ganzen auf einem Kontinuum). Es herrscht daher in diesem Punkt Uneinigkeit im Forschungsgebiet zu intuitiver Benutzung. Unabhängig von diesem Aspekt wurden die obigen Merkmale so formuliert, dass sie die vollkommene intuitive Benutzung und somit das positive Extremum des Kontinuums repräsentieren. Je weiter man sich von den aufgelisteten Ausprägungen der Merkmale entfernt, umso geringer wird die intuitive Benutzung des Systems, wobei man für eine detaillierte Interpretation die unterschiedlichen Ansichten der Forschergruppen bezüglich des Auftretens der genannten Merkmale mitberücksichtigen sollte. Zusätzlich lässt sich intuitive Benutzung von physischer Effizienz und zeitliche Effizienz im Sinne der motorischen Handlungsausführung von intuitiver Benutzung abgrenzen, wie es bereits in Teilabschnitt 2.1.1 erwähnt wurde.

2.2 Definierende Merkmale intuitiver Benutzung

Aufgrund der Vielzahl an charakteristischen Merkmalen, die im letzten Abschnitt von verschiedenen HCI-Forschergruppen mit intuitiver Benutzung in Verbindung gebracht wurden

und aufgrund des Aspekts der Uneinigkeit der Forschergruppen bezüglich des Auftretens dieser verschiedenen Merkmale (d.h. „Müssen alle Merkmale gleichzeitig auftreten, damit eine intuitive Benutzung vorliegen kann?“), wird im Folgenden erläutert, welche der genannten Merkmale den anderen Merkmalen zugrunde liegen (d.h. zentrale charakterisierende Merkmale intuitiver Benutzung, die bei jeder intuitiven Benutzung unabhängig vom Nutzungskontext immer auftreten). Die identifizierten Merkmale können auf diese Weise als Basis für eine einheitliche für die Evaluation intuitiver Benutzung geeignete Arbeitsdefinition (d.h. Messdefinition) fungieren. Dazu werden zunächst Zwei-Prozess-Theorien (siehe Evans & Stanovich, 2013; Stanovich et al., 2014) als metatheoretisches Rahmenmodell der kognitiven Informationsverarbeitung im Allgemeinen vorgestellt. Im Anschluss wird dieses Rahmenmodell zur Identifikation des zentralen subjektiven (d.h. metakognitives Gefühl von Flüssigkeit), des zentralen objektiven (d.h. mentale Beanspruchung) und des ergänzenden pragmatischen Merkmals (d.h. Effektivität) intuitiver Benutzung für die Bereitstellung dieser Messdefinition genutzt.

2.2.1 Zwei-Prozess-Theorien

Obwohl große Überschneidungen zwischen den einzelnen HCI-Definitionen intuitiver Benutzung existieren, haben die jeweiligen HCI-Forschergruppen bei einigen Merkmalen eine geteilte Meinung oder legen andere Schwerpunkte in ihrer Definition. Beispielsweise sind für die IUUI-Forschergruppe eine objektiv feststellbare effektive Benutzung und eine mental effiziente Informationsverarbeitung als Merkmale bzw. Variablen zur Beurteilung des Ausmaßes intuitiver Benutzung wichtig (siehe Hurtienne, 2011), welche von den anderen beiden Forschergruppen in dieser Form nicht explizit (d.h. nicht ausdrücklich gefordert, siehe Blackler, 2006) oder nur indirekt (d.h. nur indirekt anhand von Mühelosigkeit gefordert, siehe Ullrich, 2014) gefordert werden (siehe Blackler, 2018).

Die beiden anderen Forschergruppen (d.h. QUT- und INTUI-Forschergruppe) betonen im Gegensatz zur IUUI-Forschergruppe jedoch die Wichtigkeit der geringen wahrgenommenen expliziten Verbalisierungsfähigkeit und der wahrgenommenen zeitlichen Effizienz bei der kognitiven Informationsverarbeitung als subjektive charakterisierende Merkmale intuitiver Benutzung. Im Allgemeinen spielt die subjektive gefühlsgeleitete Seite intuitiver Benutzung für die einzelnen Forschergruppen eine unterschiedlich starke Rolle. Sie wird am stärksten von der INTUI-Forschergruppe berücksichtigt (siehe Blackler, 2018; Blackler & Popovic, 2015; Ullrich, 2014). Zusätzlich geht die INTUI-Forschergruppe als einzige davon aus, dass sich das Ausmaß an intuitiver Benutzung nicht gleichermaßen in allen gesammelten Merkmalen widerspiegeln muss, sondern vielmehr die relative Ausprägung der Merkmale zueinander das Ausmaß intuitiver Benutzung beschreibt (siehe Ullrich, 2014). Die anderen beiden Forschergruppen (d.h. QUT- und IUUI-Forschergruppe) postulieren hier hingegen (siehe Blackler, 2006; Hurtienne, 2011), dass sich das Ausmaß intuitiver Benutzung immer gleichermaßen in den, von der jeweiligen Forschergruppe angegebenen, charakteristischen Merkmalen widerspiegelt.

Derartige Unterschiede bezüglich einer intuitiven Benutzung zugeordneten Merkmale und die Frage, welche Merkmale nun ausreichen, um überhaupt von einer intuitiven Entscheidung sprechen zu können, gibt es nicht nur im HCI-Bereich, sondern auch in der allgemeinen übergeordneten Intuitionsforschung (siehe Evans & Stanovich, 2013). Dies hat eine

uneindeutige Terminologie bei der Beschreibung eines intuitiven Verarbeitungsprozesses und der damit verbundenen charakterisierenden Merkmale zur Folge (siehe Tabelle 2.2). Evans (2010) beispielsweise sieht eine zusammenfassende Beschreibung intuitiver Handlungen als „unbewusst“ sehr problematisch (d.h. eine Handlung ist dann schon intuitiv, wenn sie unbewusst ist), da viele vage Definitionen von Unbewusstsein existieren und die damit verbundenen Eigenschaften stark variieren können. Eine ausschließlich auf diesem Merkmal basierende Dichotomisierung (d.h. unbewusst vs. bewusst) kann außerdem den Eindruck vermitteln, dass intuitive Handlungen überhaupt keine bewusste Verarbeitung zulassen und eine intuitive Handlung nicht auf einem Kontinuum zwischen unbewusster und bewusster kognitiver Informationsverarbeitung variieren kann. Eine ähnliche Problematik wie sie schon bei der Definition der unbewussten Anwendung von handlungsrelevantem Vorwissen als Grundvoraussetzung intuitiver Benutzung im Teilabschnitt 2.1.1 dieser Arbeit erläutert wurde.

Das Problem für derartige Ungereimtheiten sehen eine Reihe von Intuitionsforschern außerhalb der HCI (z.B. Evans, 2011; Evans & Stanovich, 2013; B. R. Newell & Shanks, 2014; Pfister et al., 2016; Zander et al., 2016) in den verschiedenen Ansätzen bezüglich der Verortung intuitiven Handelns innerhalb des menschlichen Denkens. Die den Definitionen intuitiver Benutzung im HCI-Bereich zugrunde liegende Literatur stammt aus diversen Forschungsgebieten, in welchen die kognitive Informationsverarbeitung des Menschen üblicherweise aus zwei verschiedenen Perspektiven betrachtet wird. Entweder erfolgt die Betrachtung aus einer *Ein-Prozess-Perspektive*, welche das menschliche Denken und Problemlösen als Ergebnis eines einzigen kognitiven Prozesses sieht (z.B. Keren & Schul, 2009; Kruglanski & Gigerenzer, 2011; Mega, Gigerenzer, & Volz, 2015; Osman, 2004), oder sie erfolgt aus einer *Zwei-Prozess-Perspektive*, in welcher sie durch die Wechselwirkung von zwei voneinander klar getrennten kognitiven archetypischen Prozessen (z.B. bewusste Akkommodation vs. unbewusste Assimilation) realisiert wird (z.B. Epstein, 1994; Evans & Stanovich, 2013; Glöckner & Betsch, 2008; Klein, 1998; Reyna & Brainerd, 1995; Stanovich, 1999; Stanovich et al., 2014; V. Thompson et al., 2011; Tversky & Kahneman, 1974).

Von sogenannten *Zwei-Prozess-Theorien*, die die menschliche kognitive Informationsverarbeitung aus einer Zwei-Prozess-Perspektive betrachten, existieren seit den 70er Jahren unüberschaubar viele, weswegen im Rahmen dieser Arbeit nur allgemein von Zwei-Prozess-Theorien gesprochen wird. Diese Theorien werden daher lediglich aus einer metatheoretischen Perspektive betrachtet. Sie stammen unter anderem aus den Forschungsgebieten, die die Grundlage für die bereits vorgestellten Definitionen intuitiver Benutzung bilden. Beispielsweise gibt es solche Theorien schon lange im Bereich der Kognitionspsychologie (z.B. Shiffrin & Schneider, 1977; Sloman, 1996; Stanovich, 1999; Stanovich & West, 2000), der Sozialpsychologie (z.B. Chaiken & Trope, 1999; Serena Chen & Chaiken, 1999; Epstein, 1994) und der Entscheidungspsychologie (z.B. Hogarth, 2001; Plessner & Czenna, 2011; Tversky & Kahneman, 1974). Darüber hinaus sind einige dieser Zwei-Prozess-Theorien auf die Erklärung bestimmter Phänomene wie Problemlösen (z.B. Epstein, 1994; Sloman, 1996), Verhaltens- bzw. Einstellungsänderung (z.B. Chaiken, 1987; Serena Chen & Chaiken, 1999) und Urteils- bzw. Entscheidungsfindung (z.B. Evans, 2011; Hogarth, 2001) spezialisiert. Für ein ausführliches Review bezüglich Zwei-Prozess-Theorien wird an dieser Stelle auf Evans und Stanovich (2013), Horstmann (2012) und Pfister et al. (2016) verwiesen.

Trotz ihrer unterschiedlichen Ursprünge unterscheiden alle Zwei-Prozess-Theorien zwischen einem unbewussten (d.h. Assimilation) und einem bewussten (d.h. Akkommodation) kognitiven Informationsverarbeitungsprozess. Es existieren dabei eine Reihe von Bezeichnungen für diese beiden archetypischen Prozesse (siehe Tabelle 2.2). Da beide Verarbeitungsprozesse, wie bereits im Rahmen der Definitionen intuitiver Benutzung verdeutlicht (siehe Abschnitt 2.1), mit einer Reihe von objektiven und subjektiven Merkmalen assoziiert sind, kann dies oftmals durch ein einzelnes Begriffspaar bzw. charakterisierendes Merkmal nicht ausgedrückt werden.

Tabelle 2.2. *Auswahl an einigen alternativen Bezeichnungen für Assimilation (intuitive Verarbeitung) und Akkommodation (reflektierende Verarbeitung), wie sie in verschiedenen Zwei-Prozess Theorien zu finden sind (siehe Horstmann, 2012; Stanovich, West, & Toplak, 2014). Die Tabelle ist nicht geordnet, um keine Wertung der Bezeichnungen zu suggerieren.*

Zwei-Prozess Theorie	Typ 1 Prozess (Assimilation)	Typ 2 Prozess (Akkommodation)
Bargh und Chartrand (1999)	Automatische Verarbeitung	Bewusste Verarbeitung
Bazerman, Tenbrunsel und Wade-Benzoni (1998)	Wollendes Ich	Sollendes Ich
Bickerton (1995)	Online Denken	Offline Denken
Brainerd und Reyna (2001)	Quintessenz Verarbeitung	Analytische Verarbeitung
Serena Chen und Chaiken (1999)	Heuristische Verarbeitung	Systematische Verarbeitung
Evans (1984)	Heuristische Verarbeitung	Analytische Verarbeitung
Evans und Over (1996)	Impliziter Denkprozess	Expliziter Denkprozess
Evans und Wason (1976)	Typ 1 Prozesse	Typ 2 Prozesse
Fodor (1983)	Modulare Prozesse	Zentrale Prozesse
Gawronski und Bodenhausen (2006)	Assoziierende Prozesse	Aussagende Prozesse
Haidt (2001)	Intuitives System	Schlussfolgerndes System
Johnson-Laird (1983)	Implizite Schlussfolgerungen	Explizite Schlussfolgerungen
Kahneman und Frederick (2002)	Intuition	Schlussfolgerung
Lieberman (2003)	Reflexives System	Reflektierendes System
Loewenstein (1996)	Instinktive Faktoren	Vorlieben
Metcalfe und Mischel (1999)	Heißes System	Kühles System
D. A. Norman und Shallice (1986)	Ablaufplanung	Überwachendes Aufmerksamkeitssystem
Pollock (1996)	Schnelle und inflexible Module	Verstehen
Posner und Snyder (2004)	Automatische Aktivierung	Bewusste Verarbeitung
A. S. Reber (1989)	Implizite Kognition	Explizites Lernen
Shiffrin und Schneider (1977)	Automatische Verarbeitung	Kontrollierte Verarbeitung
Sloman (1996)	Assoziierendes System	Regelbasiertes System
E. R. Smith und DeCoster (2000)	Assoziierende Verarbeitung	Regelbasierte Verarbeitung
Stanovich (2005)	Autonomes Prozesse	Entkoppelte Simulation
Strack und Deutsch (2004)	Impulsives System	Reflektierendes System
Thaler und Shefrin (1981)	Macher	Planer
Toates (2006)	Stimulus-gebunden	Höhere Ordnung
T. D. Wilson (2004)	Adaptiv unbewusst	Bewusst
Hammond (1993)	Intuitiv	Analytisch
Epstein (1994)	Erfahrungsbasiert	Rational
Kahneman (2003)	System 1	System 2
Klaczynski (2000)	Heuristische Verarbeitung	Analytische Verarbeitung
V. Thompson, Turner und Pennycook (2011)	Heuristische Prozesse	Analytische Prozesse

Die Begriffspaare beziehen sich meistens nur auf einen Aspekt und erwecken damit den Anschein, dass sich der Prozess auf dieses Attribut (z.B. automatisch, implizit, schnell) reduzieren lässt und nur dieses Attribut ausreichend ist, weswegen sich alle mit dem Prozess assoziierten konkreten Merkmale nur schwer damit in Verbindung bringen lassen (Evans & Stanovich, 2013; Stanovich et al., 2014). Beispielsweise stand das Begriffspaar *System 1* und *System 2* aufgrund seiner uneindeutigen Terminologie schon oft zur Debatte. Die Begriffe wurden ursprünglich als ziemlich generisch von Stanovich (1999) eingeführt, um zwei

unterschiedliche Eigenschaftsprofile beschreiben zu können. Mit zunehmender Popularität der Begriffe (siehe Kahneman, 2003), entstanden zunehmend Probleme (Evans & Stanovich, 2013). Die Begriffe System 1 und System 2 könnten implizieren, dass Menschen über zwei voneinander getrennte Systeme für verschiedene Arten kognitiver Verarbeitung verfügen, die sich nicht dieselben kognitiven Ressourcen teilen und sogar auf neurologischer Ebene distinkt voneinander sind (Evans, 2008, 2011; Evans & Stanovich, 2013; Stanovich & Toplak, 2012). Um derartigen Fehlinterpretationen entgegenzuwirken, wird in der Literatur mittlerweile von *Typ 1* (d.h. vollständig unbewusst und intuitiv) und *Typ 2* (d.h. vollständig bewusst und reflektierend) Prozessen gesprochen, um die zwei kognitiven archetypischen Verarbeitungsmechanismen (d.h. unbewusste Assimilation und bewusste Akkommodation) unterscheiden zu können (Evans & Stanovich, 2013; Stanovich et al., 2014). Diese Begriffe sollen auch im weiteren Verlauf dieser Arbeit Anwendung finden, um Uneindeutigkeiten zu vermeiden.

Unabhängig davon für welche Bezeichnung sich die Autoren der jeweiligen Zwei-Prozess-Theorie entschieden haben, stimmen Zwei-Prozess-Theorien immer darin überein, dass sie jedem der zwei kognitiven archetypischen Verarbeitungsprozesse distinkte Prozesseigenschaften zuordnen (Evans & Stanovich, 2013; Pfister et al., 2016; Stanovich et al., 2014; Zander et al., 2016). Dabei assoziieren die Mehrheit der Autoren, dass eine vollkommen intuitive kognitive Informationsverarbeitung objektiv durch eine schnelle, unbewusste, automatische und parallele Verarbeitung mit einer hohen Verarbeitungskapazität unter gleichzeitig geringer mentaler Beanspruchung (d.h. hohe mentale Effizienz) gekennzeichnet ist. Eine vollkommen reflektierende kognitive Informationsverarbeitung ist hingegen objektiv durch eine langsame, kontrollierte, bewusste, mental beanspruchende (d.h. geringe mentale Effizienz) und serielle Verarbeitung mit einer geringeren Verarbeitungskapazität charakterisiert (Evans & Stanovich, 2013; Horstmann, 2012; Pfister et al., 2016).

Abhängig von der thematischen Ausrichtung der jeweiligen Zwei-Prozess-Theorie variiert die Liste der Prozesseigenschaften, welche den beiden Verarbeitungsprozessen zugeschrieben werden (Evans, 2011; Evans & Stanovich, 2013; Horstmann, 2012; Stanovich, 1999). Die vorgestellten HCI-Definitionen intuitiver Benutzung verwenden dabei unterschiedliche Zwei-Prozess-Theorien als theoretische Grundlage. So nutzt Hurtienne (2011) für seine IUUI-Definition beispielsweise die Theorie von D. A. Norman und Shallice (1986) und Blackler (2006) stützt ihre QUT-Definition auf den Überlegungen von Shiffrin und Schneider (1977). Die INTUI-Forschergruppe (z.B. Diefenbach & Ullrich, 2015; Ullrich, 2013, 2014) bezieht sich entweder auf die anderen beiden Forschergruppen oder verwendet im Gegensatz Literatur, die die kognitive Informationsverarbeitung überwiegend aus einer Ein-Prozess-Perspektive betrachtet (siehe Gigerenzer, 2007). Auf diese Perspektive wird an späterer Stelle dieses Abschnitts genauer eingegangen.

Evans und Stanovich (2013) fassen in ihrem Review-Artikel die wesentlichen Prozesseigenschaften bzw. die assoziierten Merkmale der beiden komplementären Verarbeitungsprozesse zusammen. Die Prozesseigenschaften, welche dort einem vollkommen intuitiven Verarbeitungsprozess zugeordnet werden, entsprechen überwiegend den objektiven Merkmalen, die auch im HCI-Bereich mit intuitiver Benutzung in Verbindung gebracht werden. Die Tabelle 2.3 illustriert diese Prozesseigenschaften. Beim Betrachten der Auflistung von Evans und Stanovich (2013) fällt auf, dass die beiden Informationsverarbeitungsprozesse überwiegend anhand von objektiven Merkmalen charakterisiert werden (z.B. schnelle,

geringe Beanspruchung des Arbeitsgedächtnisses), so wie sie auch explizit in den Definitionen intuitiver Benutzung der QUT- und IUUI-Forschergruppen zu finden sind (siehe Abschnitt 2.1).

Tabelle 2.3. *Typische charakterisierende Merkmale von Typ 1 (d.h. vollkommen intuitiv; Assimilation) und Typ 2 (d.h. vollkommen reflektierend; Akkommodation) Prozessen (siehe Horstmann, 2012). Die aufgrund der in Abschnitt 2.1 vorgestellten Definitionen mit intuitiver Benutzung assoziierten Merkmale sind in der Tabelle kursiv hervorgehoben.*

Typ 1 Prozess (Assimilation)	Typ 2 Prozess (Akkommodation)
schnell	*langsam*
unbewusst	*bewusst*
hohe Verarbeitungskapazität	geringe Verarbeitungskapazität
parallel	sequentiell
automatisch	kontrolliert
assoziativ	regelbasiert
geringe Beanspruchung des Arbeitsgedächtnisses	*hohe Beanspruchung des Arbeitsgedächtnisses*
kontextabhängig	abstrakt
affektiv	*nicht affektiv*
unabhängig von kognitiven Fähigkeiten	abhängig von kognitiven Fähigkeiten

Subjektive Konsequenzen intuitiver Benutzung, wie sie insbesondere von der INTUI-Forschergruppe vorgeschlagen werden, fehlen in dieser Auflistung, obwohl eine Vielzahl der aktuell bedeutsamsten Zwei-Prozess-Theorien explizit die Wichtigkeit affektiver Informationen als präattentive metakognitive Hinweise für die kognitive Informationsverarbeitung betonen, da sich daraus das bewusste subjektive Gefühl von Flüssigkeit entwickelt (z.B. Ackerman & Thompson, 2017; Hogarth, 2001; Kahneman, 2003; R. Reber et al., 2002; R. Reber et al., 2004; Reyna, 2008; V. Thompson, 2009). Da der damit verbundene präattentive affektive Richtimpuls mithilfe von Typ 1 Prozessen realisiert wird, spielen affektive Informationen für Typ 2 Prozesse keine Rolle (Ackerman & Thompson, 2017; Horstmann, 2012; V. Thompson, 2009). Dementsprechend wurde die Tabelle 2.3 von Evans und Stanovich (2013) um diese affektive Eigenschaft in Anlehnung an Kahneman (2003) und Horstmann (2012) erweitert. An dieser Stelle wird außerdem darauf hingewiesen, dass das in Teilabschnitt 2.1.1 vorgestellte Leistungskriterium *Effektivität* nicht in der Tabelle berücksichtigt ist, da sich Effektivität immer auf die ganze Handlung und nicht nur auf die kognitive Informationsverarbeitung bezieht, die mithilfe von Typ 1 und Typ 2 Prozessen realisiert wird. Trotz einer generellen Übereinstimmung bezüglich der Eigenschaften beider archetypischer Prozesse, die die jeweilige maximale Ausprägung der damit assoziierten Merkmale und somit die beiden Enden des kognitiven Kontinuums des Menschen repräsentieren, ergibt sich bei genauerem Hinsehen eine zentraler Unterschied. Dieser Unterschied betrifft die Konzeptualisierung der Interaktion zwischen den beiden Prozessen (Evans, 2007; Horstmann, 2012). Bezüglich der Interaktion beider Prozesse können laut Evans (2007) zwischen *präemptiven Theorien*, *parallel-kompetitiven Theorien* und *Default-Interventionist-Theorien* unterschieden werden.

Kennzeichnend für präemptive Theorien ist die Auffassung, dass bei der ersten Konfrontation mit einer Aufgabe eine Art der beiden kognitiven Informationsverarbeitungsprozesse (d.h. entweder unbewusste Assimilation oder bewusste Akkommodation) gewählt und anschließend für die vollständige Lösung der Aufgabe verwendet wird. Es findet nach dieser

Entscheidung keinerlei Interaktion bzw. Wechsel zwischen den beiden Prozessen statt. Laut Evans (2007) wird lediglich die Theorie von Klaczynski (2000) dieser Klasse von Theorien zugeordnet (siehe Horstmann, 2012). Dementsprechend findet laut dieser Art von Theorien intuitive Benutzung nicht auf einem Kontinuum statt, sondern wird durch einen einzigen, vollständig intuitiven Assimilationsprozess abgebildet, der über alle in der Tabelle 2.3 abgebildeten Eigenschaften verfügt.

Parallel-kompetitive Theorien (z.B. Serena Chen & Chaiken, 1999; Epstein, 1994; A. S. Reber, 1989; Shiffrin & Schneider, 1977; Sloman, 1996) postulieren, dass beide Arten von kognitiven Informationsverarbeitungsprozessen gleichzeitig aktiviert werden und es dann zu eine Art Wettstreit zwischen den beiden Prozessarten kommt, was zu im Konflikt stehenden Antworttendenzen führen kann (Horstmann, 2012). Wie präemptive Theorien unterstellen diese Art von Zwei-Prozess-Theorien ebenfalls, dass der Mensch lediglich über zwei dichotome kognitive Verarbeitungsprozesse verfügt, die eine Reihe von ebenfalls dichotomen Eigenschaften besitzen (Evans & Stanovich, 2013). Ein kognitives Kontinuum würde bezüglich dieser Art von Theorien nicht existieren, sondern die Verarbeitung ähnlich wie bei präemptiven Theorien von zwei dichotomen Prozessen (d.h. vollkommen intuitiv und vollkommen reflektierend) übernommen werden. Laut Evans und Stanovich (2013) ist das Problem bei einer solchen Auffassung, dass Typ 1 Prozesse per Definition immer schneller als Typ 2 Prozesse sind. Wenn Assimilation und Akkommodation wirklich immer gleichzeitig aktiviert werden würden, müsste der schnelle Typ 1 Prozess immer auf den langsamen Typ 2 Prozess warten, bis ein möglicher Konflikt auflösbar wäre. Ein weiteres Problem dieser Auffassung ist zusätzlich, dass der langsame Typ 2 Prozess, um analytisch arbeiten zu können, ein hohes Maß an kognitiven Ressourcen blockiert und erst wieder freigeben kann, wenn ein Konflikt aufgelöst ist. Daher ist diese Art der Interaktion für die Erklärung der kognitiven Verarbeitung ähnlich wie präemptive Theorien nur bedingt geeignet und spiegelt die menschliche Handlungsregulation nicht korrekt wider, so wie es für ein Kontinuum benötigt wird (Evans & Stanovich, 2013; Stanovich et al., 2014).

Aufgrund dieser Dichotomisierung kritisierten in der Vergangenheit neben Evans und Stanovich (2013) eine Reihe weiterer Forscher die empirische Evidenz parallel-kompetitiver Zwei-Prozess-Theorien aufgrund ihrer vagen Definition, ihrer schlechten theoretischen Ausarbeitung und der fehlenden Kohärenz und Konsistenz in den von unterschiedlichen Theorien vorgeschlagenen dichotomen Gruppen von Eigenschaften (z.B. De Neys & Glumicic, 2008; Keren & Schul, 2009; Kruglanski & Gigerenzer, 2011; Osman, 2004; Topolinski & Strack, 2009). Die Grundlage für den Beginn dieser Debatte bildet die von Stanovich (1999) erstmals veröffentlichte Eigenschaftstabelle, wo dieser, in ähnlicher Weise wie Evans und Stanovich (2013) (siehe Tabelle 2.3), die bis zum damaligen Zeitpunkt genannten Merkmale zur Charakterisierung der beiden Prozesse gegenüberstellt. Die Liste von Stanovich (1999) war, wie andere von über zwei Dutzend anderen Theoretikern stammende Listen (siehe Evans & Stanovich, 2013), nicht vorgesehen, um als strenge theoretische Aussage bezüglich erforderlicher (d.h. notwendige gleichzeitig auftretende) und damit definierender Merkmale (d.h. alleine diese Merkmale reichen aus um zwischen beiden archetypischen Prozessen unterscheiden zu können) der beiden Prozesse zu fungieren (Evans, 2011; Evans & Stanovich, 2013; Stanovich et al., 2014).

Laut der Aussage von Stanovich (1999) sollte die Liste lediglich seine Suche nach Ähnlichkeit inmitten der verschiedenen Theorien widerspiegeln. Die Liste war für ihn rein

deskriptiv und dementsprechend nicht empirisch fundiert, da es aufgrund der damaligen unsystematischen und nicht quer verwiesenen empirischen Befunde auch nicht anders möglich gewesen wäre (Evans, 2011; Evans & Stanovich, 2013; Stanovich et al., 2014). Es handelte sich bei den, dem vollständig intuitiven und vollständig reflektierenden Verarbeitungsprozess zugeordneten, charakterisierenden Merkmalen lediglich um eine zufällige empirisch schlecht fundierte persönliche Auswahl (Evans & Stanovich, 2013). Nichtsdestotrotz nutzten eine Reihe von Kritikern (z.B. Keren & Schul, 2009; Kruglanski & Gigerenzer, 2011; Osman, 2004) diesen Umstand zur Bildung eines Strohmann-Arguments (d.h. rhetorische Technik bei der die tatsächliche Auseinandersetzung mit der Gegenseite nur fingiert und stattdessen gegen einen fiktiven Strohmann argumentiert wird, welchem eine undifferenzierte Version der gegnerischen Argumentation in den Mund gelegt wird, siehe Walton, 2013) aus. Denn je länger die Liste von Merkmalen auf einer Prozess-Seite (d.h. intuitiv oder reflektierend) ist, umso einfacher lässt sich daraus ein Strohmann-Argument konstruieren. Denn falls nicht alle Merkmale von einer Seite gleichzeitig auftreten, so lässt sich auch die allgemeine Gültigkeit von parallel-kompetitiven Zwei-Prozess-Theorien stark anzweifeln (Evans, 2011; Evans & Stanovich, 2013; Stanovich et al., 2014).

Neben Osman (2004) und Keren und Schul (2009) konstruierten Kruglanski und Gigerenzer (2011) zuletzt ein solches Strohmann-Argument, indem sie in ihrer sog. „Alignment Assumption“ behaupteten, dass parallel-kompetitive Zwei-Prozess-Theorien im Allgemeinen inkorrekt sein müssen, da die postulierten Merkmale auf jeder Seite mehr oder weniger stark miteinander verbunden sind. Für die explizite Formulierung ihres Strohmann-Arguments nahmen sie die ursprünglichen sechs dichotomen Merkmale von Stanovich (1999) und interpretierten diese als definierend für die beiden kognitiven Verarbeitungsprozesse. Infolgedessen führte dies dazu, dass alle Merkmale auf einer Prozessseite gleichzeitig auftreten müssen, damit es sich um die jeweilige Art der kognitiven Informationsverarbeitung handelt. Dies würde laut Kruglanski und Gigerenzer (2011) jedoch darin resultieren, dass man bei sechs Dichotomien eine $2^6 = 64$ Zell-Matrix erhält, bei der lediglich zwei Zellen gefüllt sind, da diese Zellen die Verbindung aller sechs Dichotomien enthalten. Eine Tatsache, die empirisch von keiner Zwei-Prozess-Theorie gestützt ist. Kruglanski und Gigerenzer (2011) bezogen sich in ihrem Artikel auf keine bestimmte parallel-kompetitive Zwei-Prozess-Theorie, da auch keine Zwei-Prozess-Theorie existiert, die einen solchen Anspruch überhaupt explizit erhebt. Sie argumentierten stattdessen allgemein auf einer Metaebene gegen Zwei-Prozess-Theorien. Die Strategie von Kruglanski und Gigerenzer (2011) ist nicht verwunderlich, da solche Dichotomien lediglich als Eigenschaften in Form von Korrelaten zur Organisation der verfügbaren Literatur gedacht waren und nicht dazu verwendet werden sollten, eine spezifische theoretische Aussage bezüglich des gleichzeitigen Auftretens von Prozesseigenschaften zu treffen. Derartige Vorhersagen können mit keiner existierenden psychologischen Theorie getroffen werden, egal wie fundiert diese ist (Evans, 2011; Evans & Stanovich, 2013; Stanovich et al., 2014). Für eine detaillierte Übersicht weiterer Kritikpunkte und Strohmann-Argumente anderer Autoren (z.B. Keren & Schul, 2009; Osman, 2004) wird an dieser Stelle auf Evans und Stanovich (2013) und Horstmann (2012) verwiesen.

Kritiker parallel-kompetitiver Zwei-Prozess-Theorien wie Newstead (2000), Osman (2004) und Kruglanski und Gigerenzer (2011) sind der Auffassung, dass die kognitive Informationsverarbeitung eines Menschen durch einen einzigen Prozess erklärt werden kann (d.h. Ein-Prozess-Perspektive). Die Merkmale dieses einen Prozesses (z.B. Bewusstseinspflich-

tigkeit bei der kognitiven Verarbeitung, zeitliche Effizienz bei der kognitiven Verarbeitung) varrieren dabei entlang eines Kontinuums. Die kognitive Informationsverarbeitung findet dementsprechend nicht durch paralleles Aktivieren von dichotomen kognitiven Prozessen statt. Diese Auffassung wird auch von anderen Forschern, welche aber trotzdem am prinzipiellen Konzept einer Zwei-Prozess-Perspektive festhalten, überwiegend so akzeptiert (Evans & Stanovich, 2013). Als Reaktion auf die Kritik gegenüber parallel-kompetitiven Theorien und auf Basis des Gedankens eines kognitiven Kontinuums entwickelten sich Default-Interventionist-Theorien, die als Hybrid von Zwei-Prozess und Ein-Prozess Theorien gelten (Evans & Stanovich, 2013; B. R. Newell, Lagnado, & Shanks, 2015).

Default-Interventionist-Theorien (Evans, 2006, 2011; Glöckner & Betsch, 2008; Kahneman & Frederick, 2002; Stanovich et al., 2014; V. Thompson, 2009) postulieren, dass intuitive Verarbeitungsprozesse per Default (d.h. Voreinstellung) immer zuerst aktiviert werden und analytische Prozesse erst in einem zweiten Schritt, falls nötig, intervenieren können (Evans & Stanovich, 2013; Horstmann, 2012; Stanovich et al., 2014). Verschiedene Phänomene in der menschlichen Entscheidungsfindung bei der Handlungsregulation bzw. der damit verbundenen kognitiven Informationsverarbeitung können von dieser Klasse am besten erklärt werden. Es herrscht daher bezüglich der Gültigkeit dieser Art von Theorien in der Forschungsliteratur auch bei früheren Kritikern von parallel-kompetitiver Zwei-Prozess-Theorien die größte Einigkeit (z.B. Evans, 2007, 2009; Evans & Stanovich, 2013; Frankish, 2010; Horstmann, 2012; Kahneman & Egan, 2011; Kruglanski, 2013; Stanovich et al., 2014; V. Thompson, Evans, & Campbell, 2013).

Im Gegensatz zu den anderen beiden vorgestellten Klassen von Zwei-Prozess-Theorien unterscheiden Default-Interventionist-Theorien, wie beispielsweise die Theorien von Evans (2007, 2010) und Stanovich (1999, 2011), explizit zwischen dichotomen archetypischen Prozesstypen, welche sich per Definition qualitativ bezüglich ihrer Eigenschaften voneinander unterscheiden (d.h. Typ 1 ist vollkommen intuitiv und Typ 2 ist vollkommen reflektierend, siehe Abbildung 2.3), und der tatsächlich stattfindenden kognitiven Informationsverarbeitung während der Handlungsregulation, die auf einem Kontinuum variiert und somit für eine erfolgreiche Handlungsregulation beide Prozessarten benötigt. Das kognitive Kontinuum ergibt sich aufgrund der Möglichkeit von Typ 2 Prozessen, falls dies durch die Situation erforderlich scheint, bewusst zu intervenieren und damit die kognitive Informationsverarbeitung zu übernehmen. Bei einer intuitiven Handlung wird diese überwiegend von Typ 1 Prozessen reguliert, wobei eine reflektierende Handlung überwiegend von Typ 2 Prozessen reguliert wird (Evans, 2010; Evans & Stanovich, 2013; Stanovich et al., 2014). Um das dynamische Zusammenspiel von Typ 1 und Typ 2 Prozessen während der kognitiven Informationsverarbeitung im Rahmen der Handlungsregulation zu beschreiben, reicht die dichotome Unterscheidung von Typ 1 und Typ 2 Prozessen von parallel-kompetitiven Theorien nicht mehr aus, weswegen zusätzlich zu diesen Prozessen übergeordnete metakognitive Prozesse existieren müssen.

Default-Interventionist-Theorien gehen deswegen generell von einer dreigeteilten kognitiven Bewusstseinshierarchie aus (Ackerman & Thompson, 2017; Evans, 2009; Evans & Stanovich, 2013; Stanovich et al., 2014). Im Drei-Instanzen-Modell des Geistes (d.h. tripartite model of the mind) bilden unbewusste Typ 1 Prozesse den autonomen Geist und werden auch als autonome Menge von Systemen (d.h. autonomous set of systems) bezeichnet (Stanovich et al., 2014). Innerhalb von bewussten Typ 2 Prozessen wird zwischen einem

algorithmischen Geist und einem reflektierenden Geist unterschieden. Die Funktionen der einzelnen Instanzen sind in Abbildung 2.4 dargestellt und mit den Regulationsebenen aus der Handlungsregulationstheorie von Hacker (1986) verknüpft. Sie werden im folgenden Abschnitt nacheinander erläutert und mit den in Abschnitt 2.1 dargestellten Überlegungen in Verbindung gebracht.

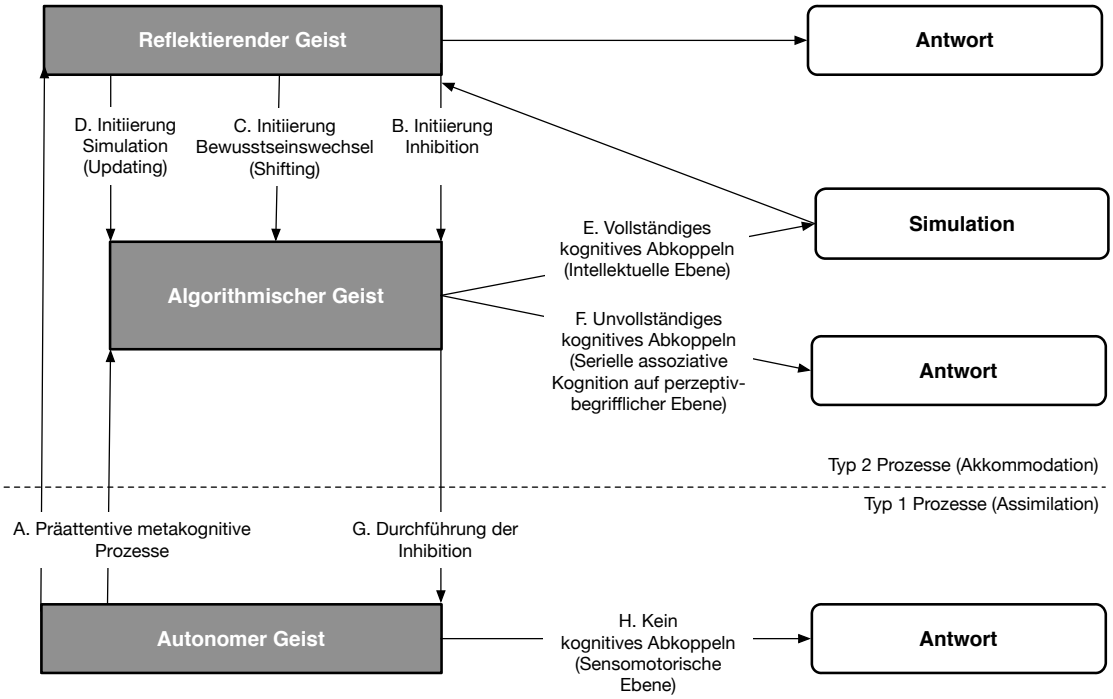


Abbildung 2.4. Drei-Instanzen-Modell des Geistes nach Stanovich, West und Toplak (2014), ergänzt mit entsprechenden Regulationsebenen aus der Handlungsregulationstheorie von Hacker (1986).

Die kognitive Informationsverarbeitung beginnt laut dem Drei-Instanzen-Modell des Geistes damit, dass die oberen beiden Instanzen vom autonomen Geist bei einem zu lösenden Handlungsproblem zunächst Input bekommen, der mithilfe sog. präattentiver metakognitiver Prozesse (siehe Pfeil A in Abbildung 2.4) erzeugt wird (Stanovich et al., 2014). Diese präattentiven metakognitiven Prozesse (siehe Abbildung 2.5) beinhalten unter anderem das bereits angesprochene metakognitive Gefühl von Flüssigkeit (d.h. subjektive Flüssigkeit oder Gefühl von Richtigkeit: Eingriff von Typ 2 Prozessen, wenn der Abruf durch Typ 1 Prozesse als nicht flüssig wahrgenommen wird), welches von Default-Interventionist-Theorien als Auslöser für die Intervention von bewussten Typ 2 Prozessen verantwortlich gemacht wird (Ackerman & Thompson, 2017; Barrouillet, Portrat, & Camos, 2011; Evans & Stanovich, 2013; Stanovich et al., 2014; V. Thompson, 2009). Default-Interventionist-Theorien postulieren hier, dass bei der Assimilation zur Lösung eines Handlungsproblems (d.h. Generierung einer Antwort oder Entscheidung) vom autonomen Geist zwei verschiedene Outputs erzeugt werden, die dann als Input sowohl für den reflektierenden, als auch für den autonomen Geist fungieren. Zum einen den *Inhalt der initialen Handlungstendenz* basierend auf dem mentalen Modell (d.h. Antwort, Default-Wert, Heuristik) und zum anderen ein *begleitendes metakognitives Gefühl von Flüssigkeit* beim Abruf dieser Hand-

lungstendenz (Ackerman & Thompson, 2017; Evans, 2010; Evans & Stanovich, 2013; R. Reber et al., 2002; R. Reber et al., 2004; Simmons & Nelson, 2006; Stanovich et al., 2014; V. Thompson, 2009; V. Thompson et al., 2011).

Die Generierung dieser beiden Outputs des autonomen Geistes kann dementsprechend als Abruf von impliziten Langzeitgedächtnisinhalten verstanden werden. Dieser Output enthält neben dem eigentlichen inhaltlichen Vorschlag zur Lösung des Handlungsproblems auch eine affektive Meta-Komponente (d.h. subjektives Gefühl von Flüssigkeit als präattentiver affektiver Richtimpuls), die den oberen beiden Instanzen signalisiert, inwiefern der Langzeitgedächtnisinhalt für die Bewältigung der aktuellen Handlungssituation geeignet (d.h. handlungsrelevantes Vorwissen ist richtig und differenziert genug) und damit auch zufriedenstellend ist. Fällt dieses metakognitive Gefühl von Flüssigkeit stark aus (d.h. hohe subjektive Flüssigkeit), war die Assimilation der Gedächtnisinhalte erfolgreich und der autonome Geist kann das Problem auf Basis eines bereits verfügbaren mentalen Modells alleine lösen (d.h. Vorwissen war richtig und differenziert genug). Bei dieser intuitiven Entscheidung wird die initiale Handlungstendenz des autonomen Geistes infolgedessen akzeptiert (siehe Abbildung 2.5). Ein schwaches Gefühl von subjektiver Flüssigkeit signalisiert den beiden oberen Instanzen hingegen die Notwendigkeit zur Intervention und zum Einsatz der Akkommodation, um für die Lösung des Handlungsproblems ein entsprechendes mentales Modell (d.h. ein richtiges und differenzierteres Modell) bereitstellen zu können (Fazendeiro, Winkielman, Luo, & Lorah, 2005; R. Reber et al., 2002; R. Reber et al., 2004; Stanovich et al., 2014; V. Thompson, 2009; V. Thompson et al., 2011; Zelazo, Moscovitch, & Thompson, 2007). In diesem Fall wird die initiale Handlungstendenz des autonomen Geistes infolgedessen überschrieben.

Eine Reihe von Forschungsarbeiten ist der Auffassung (z.B. Ackerman & Thompson, 2017; Price & Norman, 2008; R. Reber et al., 2002; R. Reber et al., 2004; Topolinski, 2011), dass dieses metakognitive bewusste Gefühl von Flüssigkeit (d.h. subjektive Flüssigkeit) maßgeblich aus der objektiven unbewussten Flüssigkeit beim impliziten Abruf von Langzeitgedächtnisinhalten an der Grenze zum menschlichen Bewusstsein entsteht (siehe Abbildung 2.5), indem dieses Gefühl zusätzlich affektiv signalisiert, inwiefern ein Langzeitgedächtnisinhalt bereits erfolgreich hervorgerufen wurde oder künftig hervorgerufen werden kann (R. Reber et al., 2004; V. Thompson, 2009; V. Thompson et al., 2011; Whittlesea & Leboe, 2003). Dieses Gefühl von Flüssigkeit kann dabei sogar die Illusion vermitteln, dass dieser Inhalt bereits schon einmal erlebt wurde, auch wenn dies nicht der Fall ist (Jacoby & Whitehouse, 1989). Es beschreibt dadurch kondensiert die subjektive Mühelosigkeit oder Einfachheit, mit der die initiale intuitive Antworttendenz einem unbewusst in den Sinn kommt (Alter & Oppenheimer, 2009; R. Reber et al., 2002; R. Reber et al., 2004; Simmons & Nelson, 2006; V. Thompson, 2009; V. Thompson & Morsanyi, 2012).

Im Rahmen der Forschungsliteratur wurde Flüssigkeit nicht nur aus der Perspektive der kognitiven Informationsverarbeitung, sondern auch aus einer Wahrnehmungsperspektive betrachtet (Alter & Oppenheimer, 2009; Alter, Oppenheimer, Epley, & Eyre, 2007). So argumentieren Alter et al. (2007), dass, wenn Menschen Probleme schwer wahrnehmen können, sie diese folglich auch nicht mithilfe ihrer unbewussten Typ 1 Prozesse verarbeiten können (d.h. es kann keine Assimilation erfolgen) und stattdessen auf reflektierende, analytische Typ 2 Prozesse (d.h. Akkommodation) zurückgreifen müssen. Diese Art der Flüssigkeit wird von den Autoren als Wahrnehmungsflüssigkeit (d.h. peceptual fluency)

bezeichnet, die zusammen mit der Abrufflüssigkeit (d.h. answer fluency, retrieval fluency) die generelle kognitive Informationsverarbeitungsflüssigkeit bestimmt (d.h. processing fluency) (Alter & Oppenheimer, 2009; R. Reber et al., 2002; R. Reber & Schwarz, 1999; R. Reber et al., 2004). Obwohl Wissenschaftler noch weitere Flüssigkeitsarten unterscheiden (z.B. motor fluency, siehe Casasanto & Chrysikou, 2011), konnte bisher empirisch noch kein Unterschied zwischen den einzelnen Flüssigkeitsarten festgestellt werden, da diese ähnliche Effekte auf die menschliche Entscheidungsfindung bei Handlungen ausüben und die Konstrukte daher schwer voneinander trennbar sind (Graf, Mayer, & Landwehr, 2018; Schwarz, 2004, 2015; Winkielman, Schwarz, Fazendeiro, Reber et al., 2003). Im Rahmen dieser Arbeit wird deswegen allgemein von Flüssigkeit gesprochen und für eine detaillierte Auseinandersetzung mit verschiedenen Flüssigkeitsarten auf Graf et al. (2018) oder Alter und Oppenheimer (2009) verwiesen.

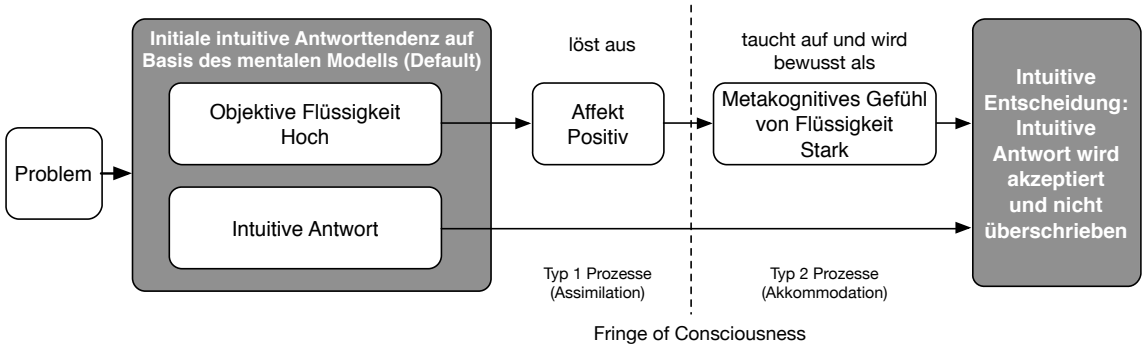


Abbildung 2.5. Präattentive metakognitive Prozesse bei intuitiver Benutzung in Anlehnung an das Doppelantwort-Paradigma (siehe Ackerman & Thompson, 2017; V. Thompson, Turner, & Pennycook, 2011) und das Flüssigkeits-Affekt-Modell der Intuition (siehe Topolinski, 2011; Topolinski & Strack, 2009).

Flüssigkeit kann demnach als die mentale und zeitliche Effizienz bei der effektiven kognitiven Informationsverarbeitung verstanden werden (objektiv und subjektiv erfahren), die bei der Wahrnehmung (z.B. R. Reber et al., 2004), der semantischen Verarbeitung (Whittlesea, 1993), dem Abruf von Gedächtnisinhalten (z.B. Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991), der motorischen Handlungsausführung und der sensomotorischen Integration (Topolinski & Sparenberg, 2012; Topolinski & Strack, 2010) stattfindet (Topolinski, 2011). In Ergänzung zur Arbeit von V. Thompson (2009) sind Alter und Oppenheimer (2009) ferner der Auffassung, dass nicht nur die Flüssigkeit egal welcher Art, sondern auch das Nichtvorhandensein dieser Flüssigkeit als metakognitiver Hinweis fungieren kann, um auf diese Weise zu signalisieren, dass eine bewusste analytische Verarbeitung benötigt wird.

Das objektive Gefühl von Flüssigkeit, das bei der kognitiven Informationsverarbeitung durch den autonomen Geist mithilfe von Typ 1 Prozessen erzeugt wird, drückt sich beim Handelnden in Form eines unbewussten affektiven Richtimpulses aus (siehe Ackerman & Thompson, 2017; Harmon-Jones & Allen, 2001; R. Reber et al., 2004; V. Thompson, 2009; V. Thompson et al., 2011; V. Thompson, Turner, Pennycook et al., 2013; Topolinski, 2011; Topolinski & Strack, 2009; Winkielman & Cacioppo, 2001). Eine hohe objektive Flüssigkeit liefert einen metakognitiven Hinweis, dass der Gedächtnisinhalt bereits schon öfter

erfolgreich abgerufen und zur Lösung eines ähnlichen Handlungsproblems genutzt wurde und somit auch sehr wahrscheinlich für den aktuellen Handlungskontext korrekt ist. Obwohl der durch die objektive Flüssigkeit hervorgerufene affektive Richtimpuls unbewusst bleibt (siehe Abbildung 2.5), kann er sich in Abhängigkeit der Stärke der objektiven Flüssigkeit bei der Assimilation in Form von diversen metakognitiven Gefühlen der Grenze zum menschlichen Bewusstsein (d.h. Fringe of Consciousness) manifestieren (siehe Kahneman & Frederick, 2002; R. Reber et al., 2004; Schwarz, 2002; Simmons & Nelson, 2006; V. Thompson, 2009; V. Thompson & Morsanyi, 2012; V. Thompson, Turner, Pennycook et al., 2013; Topolinski, 2011) und neben dem eigentlichen Inhalt selbst als Information für Entscheidungen dienen (z.B. Clore & Huntsinger, 2007; Price & Norman, 2008; Schwarz & Clore, 1983).

Alle diese Gefühle kondensieren sich dann letztendlich zu einem umfassenden Gefühl von Flüssigkeit oder Richtigkeit (Ackerman & Thompson, 2017; Price & Norman, 2008; R. Reber et al., 2004; V. Thompson, 2009; V. Thompson et al., 2011; Topolinski, 2011), was sich objektiv in einer geringen expliziten Verbalisierbarkeit bei einer mental und zeitlich effizienten kognitiven Informationsverarbeitung beim Handelnden erkennbar macht (Forster, Leder, & Ansorge, 2013, 2016; R. Reber et al., 2004). Im Rahmen der HCI-Definitionen intuitiver Benutzung spiegelt sich dieses umfassende Gefühl von Flüssigkeit dann in Konsequenz als subjektiver Zufriedenheitsaspekt in Form wahrgenommener geringer mentaler Beanspruchung bei der kognitiven Informationsverarbeitung, wahrgenommener geringer zeitlicher Beanspruchung bei der kognitiven Informationsverarbeitung und wahrgenommener geringer expliziter Verbalisierbarkeit der kognitiven Informationsverarbeitung wider (siehe Teilabschnitt 2.1.3).

Die objektive Flüssigkeit bei der Assimilation manifestiert sich im Laufe des oben beschriebenen Prozesses unter anderem auch in ein *Gefühl der Vertrautheit*, welches dem Handelnden einen Hinweis liefert, dass der gerade abgerufene Gedächtnisinhalt der gesuchte Inhalt ist (Ackerman & Thompson, 2017; Metcalfe, Schwartz, & Joaquim, 1993; Price & Norman, 2008; Schwarz, 2004). Die Stärke dieses Gefühls ist von der Höhe der objektiven Flüssigkeit abhängig (Graf et al., 2018; Schwarz, 2004; Schwarz, Sanna, Skurnik, & Yoon, 2007). Im Rahmen der HCI-Definitionen intuitiver Benutzung spiegelt sich das Gefühl der Vertrautheit als Zufriedenheitsaspekt in Form subjektiv hoch wahrgenommener Vertrautheit wider (siehe Teilabschnitt 2.1.3).

Darüber hinaus führen noch weitere durch objektive Flüssigkeit hervorgerufene Gefühle wie das *Gefühl des Irrtums* oder das *Gefühl der Kenntnis* zu einem umfassenden metakognitiven Gefühl von Flüssigkeit (Ackerman & Thompson, 2015, 2017; Gangemi, Bourgeois-Gironde, & Mancini, 2015; Koriat, 2000). Ersteres beschreibt dabei die subjektive Erfahrung, dass etwas schief gegangen ist und man einen Fehler gemacht hat (Ackerman & Thompson, 2017; Cruz, Arango-Muñoz, & Volz, 2016; De Neys, Cromheeke, & Osman, 2011; Gangemi et al., 2015). Je höher die Flüssigkeit ist, umso schwächer fällt dieses Gefühl aus (Gangemi et al., 2015). Im Rahmen der HCI-Definitionen wird dieser Zufriedenheitsaspekt als subjektiv wahrgenommene niedrige Fehlerrate und hohe wahrgenommene Zielerreichung bei der Definition intuitiver Benutzung berücksichtigt (siehe Teilabschnitt 2.1.3).

Das Gefühl der Kenntnis liefert dem Handelnden eine Beurteilung darüber, wie wahrscheinlich ein bisher noch nicht abgerufener Gedächtnisinhalt richtig erkannt wird (J. T.

Hart, 1965; Koriat, 2000; Metcalfe & Wiebe, 1987; Singer & Tiede, 2008). Laut Price und Norman (2008) lässt sich dieses Gefühl auch als „Ich-habe-etwas-auf-der-Zunge-Gefühl“ (A. S. Brown, 1991) beschreiben. Dabei findet auch eine *Beurteilung des Lernens* statt, welche dem Menschen vermittelt, wie wahrscheinlich ein gerade gelernter Inhalt zu späterer Zeit korrekt abrufbar ist (Veenman, Van Hout-Wolters, & Afflerbach, 2006). Je höher die objektive Flüssigkeit der Assimilation, umso besser ist auch das Gefühl der Kenntnis (Koriat, 1993, 2000) und die Beurteilung des Lernens (Matvey, Dunlosky, & Guttentag, 2001; Mueller, Tauber, & Dunlosky, 2013). Im Rahmen der HCI-Definitionen intuitiver Benutzung spiegeln sich das Gefühl der Kenntnis und die Beurteilung des Lernens in Form eines geringen subjektiv wahrgenommenen Lernaufwands als Maß der Zufriedenheit wider (siehe Teilabschnitt 2.1.3).

Diese und weitere Gefühle fallen umso stärker aus, je intensiver sich die Menschen zuvor mit dem gesuchten Gedächtnisinhalten beschäftigt haben und dementsprechend über ein richtiges und differenziertes mentales Modell zur Assimilation verfügen, auf das mit hoher objektiver Flüssigkeit zugegriffen werden kann (Ackerman & Thompson, 2017; Nelson, Leonesio, Shimamura, Landwehr, & Narens, 1982; V. Thompson & Morsanyi, 2012). Das aus all diesen Gefühlen resultierende, umfassende metakognitive Gefühl von Flüssigkeit zeigt sich in den in Abschnitt 2.1 vorgestellten HCI-Definitionen entweder (1) direkt als metakognitives Gefühl (siehe O’Brien et al., 2008), (2) als zufriedenstellende Benutzung mit allen damit zusammenhängenden Konsequenzen intuitiver Benutzung (z.B. Vertrautheit, Lernaufwand, siehe Naumann & Hurtienne, 2010) oder (3) als aus einem Bauchgefühl resultierende magische Erfahrung (siehe Diefenbach & Ullrich, 2015) (siehe Abbildung 2.5). Für einen vollständigen Überblick aller aktuell bekannten Gefühle, die aus objektiver Flüssigkeit resultieren und sich zusammen sich in einem Gefühl von Flüssigkeit vereinen, wird auf Ackerman und Thompson (2017) verwiesen.

In Abhängigkeit der Stärke des Gefühls subjektiver Flüssigkeit (siehe Pfeil A in Abbildung 2.4) kann der reflektierende Geist im Drei-Instanzen-Modell des Geistes als metakognitiver Überwachungsprozesses anhand der damit verbundenen präattentiven metakognitiven Prozesse feststellen, inwiefern die Antwort des autonomen Geistes zur Lösung des aktuellen Handlungsproblems (z.B. Rotation eines Objekts in einem CAD-Programm) zufriedenstellend und damit das zugrunde gelegte Vorwissen richtig und differenziert für die Handlungsregulation ausreichend ist (Stanovich et al., 2014; V. Thompson, 2009; V. Thompson et al., 2011). Ein starkes Gefühl von Flüssigkeit signalisiert dem reflektierenden Geist, dass das Handlungsproblem durch Assimilation vom autonomen Geist alleine *zufriedenstellend* lösbar ist (d.h. ein Nutzer besitzt für die Lösung der Rotationsaufgabe richtig und differenziertes Vorwissen). Es ist daher keine mentale Beanspruchung verursachende analytische Verarbeitung (d.h. Akkommodation) nötig (d.h. die initiale intuitive Antworttendenz des autonomen Geistes wird akzeptiert). Ein schwaches Gefühl von Flüssigkeit (d.h. das Vorwissen des Nutzers ist für die Lösung der Rotationsaufgabe noch nicht richtig und differenziert genug) fungiert hingegen als Auslöser für den Eingriff von bewusster, mentaler Beanspruchung verursachender Akkommodation.

Um bei einem schwachen Gefühl die Antwort des autonomen Geistes zu überschreiben und den autonomen Geist damit „offline“ zu nehmen, muss der reflektierend rationale Geist zunächst den algorithmischen Geist anweisen, die Handlungstendenz des autonomen Geistes zu überschreiben. Diese Fähigkeit wird auch als *kognitives Abkoppeln* bezeichnet, da sie

Menschen ermöglicht, mentale Simulationen zur Repräsentation diverser Entscheidungsmöglichkeiten zu konstruieren und zu koordinieren (Evans, 2007, 2010; Evans & Stanovich, 2013; Stanovich et al., 2014). Um ein solches kognitives Abkoppeln leisten zu können, muss der reflektierende Geist die drei exekutiven Grundfunktionen (d.h. Shifting, Updating, Inhibition) zunächst initiieren und diese dann vom algorithmischen Geist ausführen lassen (siehe Pfeil B bis D in Abbildung 2.4).

Damit die Antwort des autonomen Geistes überhaupt vorgenommen werden kann, muss zum einen von der unbewussten Assimilation zur bewussten Akkommodation gewechselt, und deswegen die kognitive Informationsverarbeitung fortan von Typ 2 Prozessen übernommen werden, was immer vom reflektierenden Geist initiiert werden muss (siehe Pfeil C in Abbildung 2.4) und eine Änderung des kognitiven Bewusstseins nach sich zieht (d.h. exekutive Funktion *Shifting*). Zum anderen muss mithilfe dieser Typ 2 Prozesse ein neues mentales Modell aufgebaut und auf Basis der explizit abgerufenen Langzeitgedächtnisinhalte eine Simulation vom reflektierenden Geist initiiert werden, um verschiedene Antwortmöglichkeiten für das zugrunde liegende Handlungsproblem (z.B. Rotationsaufgabe) durchspielen zu können (siehe Pfeil D in Abbildung 2.4). Diese Aktion erfordert es wiederum den Inhalt des Arbeitsgedächtnisses entsprechend zu modifizieren und zu aktualisieren (d.h. exekutive Funktion *Updating*). Schließlich kann die initiale Handlungstendenz des autonomen Geistes mithilfe der exekutiven Funktion *Inhibition* überschrieben und damit unterdrückt werden (siehe Pfeil B in Abbildung 2.4).

Der algorithmische Geist führt zur Lösung des Handlungsproblems das kognitive Abkoppeln im Anschluss entweder (1) *vollständig* (d.h. bewusste Akkommodation), (2) *teilweise* (d.h. dynamischer Wechsel zwischen unbewusster Assimilation und bewusster Akkommodation an der Grenze zum menschlichen Bewusstsein) oder (3) *gar nicht* (d.h. unbewusste Assimilation) durch. Die erste Art des kognitiven Abkoppelns (siehe Pfeil E in Abbildung 2.4) impliziert die vollständige explizite kognitive Simulation und den vollständigen Aufbau eines, für das aktuelle Handlungsproblem ausreichend richtigen und differenzierten, mentalen Modells (d.h. ein Nutzer besitzt keine Vorerfahrung mit der Rotationsaufgabe) durch Akkommodation (Stanovich, 2009, 2011; Stanovich et al., 2014). Diese Form des kognitiven Abkoppelns (siehe Abbildung 2.4) entspricht im Rahmen der Handlungsregulation (siehe Frese & Zapf, 1994; Hacker & Sachse, 2013) daher der intellektuellen Ebene (siehe Abschnitt 2.1).

Die zweite Art, welche auch als *serielle assoziative Kognition* bezeichnet wird (siehe Pfeil F der Abbildung 2.4), impliziert eine nur teilweise durchgeführte explizite kognitive Simulation, da für das Handlungsproblem prinzipiell ein mentales Modell vorliegt, um dieses zu lösen (d.h. ein Nutzer besitzt zwar Vorwissen mit der Rotationsaufgabe, welches aber noch angepasst werden muss, weil der Nutzer das Vorwissen vielleicht nur mit einem anderem CAD-Programm gesammelt hat, in dem die Benennung der Interaktionselemente beispielsweise anders war). Dieses muss jedoch für die jeweilige Situation noch an einigen Stellen bewusst angepasst werden. Die Richtigkeit und Differenziertheit des Modells reicht für eine unbewusste Assimilation zwar nicht vollständig aus, aber es werden im Gegensatz zur ersten Art des kognitiven Abkoppelns weniger bewusste Akkommodationsprozesse benötigt (Stanovich, 2009, 2011; Stanovich et al., 2014). Im Rahmen der Handlungsregulation (siehe Frese & Zapf, 1994; Hacker & Sachse, 2013) entspricht diese Art des kognitiven

Abkoppelns daher der perzeptiv-begrifflichen Ebene (siehe Abschnitt 2.1) und somit einer Verarbeitung an der Grenze zum menschlichen Bewusstsein (siehe Abbildung 2.4).

Wie zu Beginn dieses Abschnitts erwähnt, kann auf ein kognitives Abkoppeln auch gänzlich verzichtet und stattdessen die initiale Handlungstendenz des autonomen Geistes akzeptiert werden, was somit prinzipiell die dritte Möglichkeit des kognitiven Abkoppelns darstellt (siehe Pfeil H in Abbildung 2.4). Die Handlungstendenz des autonomen Geistes wird hierbei akzeptiert und nicht überschrieben, da das Handlungsproblem auf Basis des verfügbaren mentalen Modells (d.h. ein Nutzer besitzt ein für die Lösung der Rotationsaufgabe ausreichend richtiges und differenziertes mentales Modell) durch unbewusste Assimilation ohne zusätzliche bewusste Akkommodation bzw. mental beanspruchendes kognitives Abkoppeln lösbar ist. Im Rahmen der Handlungsregulation (siehe Frese & Zapf, 1994; Hacker & Sachse, 2013) entspricht diese Art des kognitiven Abkoppelns (siehe Abbildung 2.4) daher der sensomotorischen Ebene (siehe Abschnitt 2.1). Für mehr Details bezüglich des kognitiven Abkoppelns, wie es durch das Drei-Instanzen-Modell des Geistes propagiert wird, wird auf Stanovich et al. (2014) verwiesen.

Das Ausmaß des, für die Lösung eines Handlungsproblems benötigten, kognitiven Abkoppelns lässt sich wegen der daran beteiligten exekutiven Grundfunktionen (d.h. Inhibition, Updating und Shifting), welche wiederum allen anderen höheren exekutiven Funktionen implizit zugrunde liegen (Jurado & Rosselli, 2007; A. Miyake et al., 2000), an der auftretenden mentalen Beanspruchung des Arbeitsgedächtnisses bei der Handlungsregulation ablesen (Evans & Stanovich, 2013; Stanovich et al., 2014). Je stärker ein kognitives Abkoppeln für die Lösung eines Problems erforderlich ist, umso höher ist auch die Beanspruchung des Arbeitsgedächtnisses in der Konsequenz. Aus diesem Grund postulieren Default-Interventionist-Theorien, dass alleine die mentale Beanspruchung des Arbeitsgedächtnisses, welche durch das kognitive Abkoppeln entsteht, als dichotomes objektives Merkmal zur Unterscheidung zwischen Typ 1 und Typ 2 Prozessen ausreicht und sich daran die Beteiligung von Typ 1 und Typ 2 Prozessen bei der Lösung aller, während einer Handlung zu lösender Probleme ablesen lässt (Evans, 2011; Stanovich, 2011).

Da die Inhibition als exekutive Grundfunktion selbst die Kommunikation zwischen Typ 2 und Typ 1 Prozessen markiert (siehe Pfeil G der Abbildung 2.4) können Inhibitionsprozesse die gesamte mentale Beanspruchung bzw. mentale Effizienz während einer Handlung am besten widerspiegeln. Dies wurde bereits mehrfach empirisch nachgewiesen (z.B. Bailey & Iqbal, 2008; J. D. Cohen et al., 1997; A. Miyake et al., 2000; Rubinstein, Meyer, & Evans, 2001). Wie Abbildung 2.4 verdeutlichen soll, muss der algorithmische Geist alle vorgestellten Grundfunktionen ausführen, um die initiale Handlungstendenz des autonomen Geistes letztlich überschreiben und unterdrücken zu können. Je stärker dabei eine Handlungstendenz unterdrückt werden muss, umso mehr Gedächtnisinhalte müssen für eine erfolgreiche Handlungsregulation auch verändert (d.h. Updating) und der bewussten Verarbeitung zugänglich gemacht werden (d.h. Shifting), was Inhibitionsprozesse als gute Indikatoren für die Beanspruchung des Arbeitsgedächtnisses prädestiniert (Evans & Stanovich, 2013; Stanovich et al., 2014).

Auf Basis dieser theoretischen Grundlage lässt sich abschließend schlussfolgern, dass die anderen, im Rahmen von anderen Zwei-Prozess-Theorien genannten, objektiven Merkmale der beiden archetypischen Prozessarten lediglich als Korrelate anzusehen sind. Ihnen

liegt immer die Beanspruchung des Arbeitsgedächtnisses und damit auch die mentale Effizienz als zentrale Eigenschaft implizit zugrunde und bestimmt daher ihre Ausprägung auf einem Kontinuum (Evans & Stanovich, 2013; Stanovich et al., 2014). Beispielsweise ist die objektiv messbare Schnelligkeit bzw. zeitliche Effizienz bei der kognitiven Informationsverarbeitung nur als Merkmal vorhanden, wenn überhaupt eine kognitive Informationsverarbeitung und daher auch ein kognitives Abkoppeln stattfindet. Die Regulation der kognitiven Informationsverarbeitung auf der Basis des metakognitiven Gefühls von Flüssigkeit verursacht demzufolge stetig mentale Beanspruchung und erhöht damit die verbundene Abhängigkeit von bewussten Typ 2 Prozessen (Ackerman & Thompson, 2017; Markovits, Thompson, & Brisson, 2015; V. Thompson, 2009; V. Thompson et al., 2011).

Findet eine intuitive Handlung statt, signalisiert der autonome Geist ein hohes Maß an Flüssigkeit bei den damit verbundenen Handlungsproblemen, was den Bedarf an kognitivem Abkoppeln und der damit verbundenen mentalen Beanspruchung möglichst gering hält. Nur in unvorhergesehenen neuen Situationen (z.B. Planung einer neuen Handlung und damit verbundenen Zielen, Reaktion auf unerwartetes Feedback vom System) muss ein kognitives Abkoppeln zum Einsatz kommen. Eine nicht intuitive Handlung ist im Gegensatz von einem schwachen Gefühl von Flüssigkeit bei den damit verbundenen Handlungsproblemen geprägt und der Bedarf an kognitiver Abkopplung ist entsprechend höher und mit einer erhöhten mentalen Beanspruchung bzw. einer geringeren mentalen Effizienz verbunden. Basierend auf dem Gefühl von Flüssigkeit ist eine Handlung auf einem Kontinuum mehr oder weniger intuitiv, was sich nicht nur objektiv, in der damit verbundenen mentalen Beanspruchung zeigt, sondern auch subjektiv anhand der damit verbundenen Gefühle, die sich letztendlich in einem umfassenden Gefühl von Flüssigkeit oder einem Gefühl von Richtigkeit vereinen, so wie sie bereits als spezielle subjektive Maße der Zufriedenheit intuitiver Benutzung im HCI-Bereich berücksichtigt sind (siehe Abschnitt 2.1).

2.2.2 Arbeitsdefinition intuitiver Benutzung

Mithilfe von Default-Interventionist-Theorien, die postulieren, dass sich eine intuitive Handlung in ihrem Ausmaß objektiv anhand der mentalen Beanspruchung des Arbeitsgedächtnisses und subjektiv anhand des Gefühls von Flüssigkeit erkennen lässt, können auch die von der HCI in Abschnitt 2.1 als Vorbedingung oder Konsequenzen intuitiver Benutzung vorgeschlagenen charakterisierenden Merkmale intuitiver Benutzung bewertet, die Uneinigkeit bezüglich des Auftretens dieser Merkmale zwischen den Forschergruppen aufgelöst und eine Arbeitsdefinition für diese Arbeit abgeleitet werden, die sowohl für die formative als auch für die summative Evaluation intuitiver Benutzung geeignet ist. Eine derartige Bewertung und Ableitung einer neuen Messdefinition intuitiver Benutzung bringt im Vergleich zu bestehenden HCI-Definitionen eine Reihe von Vorteilen.

Durch die Konzentration auf die Beanspruchung des Arbeitsgedächtnisses bzw. die mentale Effizienz als objektives definierendes zentrales Merkmal und das Gefühl von Flüssigkeit als definierendes subjektives zentrales Merkmal lässt sich klären, ob sich das Ausmaß intuitiver Benutzung gleichzeitig in allen der von der jeweiligen Forschergruppe vorgeschlagenen Merkmalen und in gleicher Ausprägung (IUUI- und QUT-Forschergruppe) oder sie sich eher in den von der jeweiligen Forschergruppe vorgeschlagenen Merkmalen und in jeweils

unterschiedlicher Ausprägung auf einem Kontinuum zeigt (INTUI-Forschergruppe). Anhand der klaren Abgrenzung definierender zentraler Merkmale und Korrelate intuitiver Benutzung soll deutlich gemacht werden, welche Merkmale immer gleichermaßen bei einer intuitiven Benutzung vorliegen und bei welchen, von den Forschergruppen genannten, Merkmalen, es sich nur um Korrelate handelt, die in Abhängigkeit des Nutzungskontextes ebenfalls in unterschiedlicher Ausprägung beobachtet werden können (siehe Abbildung 2.4).

Tabelle 2.4. *Definierende Merkmale und Korrelate intuitiver Benutzung unter Berücksichtigung von Default-Interventionist-Theorien.*

Definierende Merkmale intuitiver Benutzung	Typische Korrelate intuitiver Benutzung
Mental effiziente kognitive Informationsverarbeitung	Überwiegend unbewusster kognitiver Informationsverarbeitungsprozess auf Basis handlungsrelevanten Vorwissens
	Zeitlich effiziente kognitive Informationsverarbeitung
	Geringe explizite Verbalisierbarkeit der kognitiven Informationsverarbeitung
Effektive Benutzung	
Zufriedenstellende Benutzung durch starkes metakognitives Gefühl von Flüssigkeit	Wahrgenommene geringe mentale Beanspruchung der kognitiven Informationsverarbeitung
	Wahrgenommene hohe Zielerreichung
	Wahrgenommene geringe zeitliche Beanspruchung der kognitiven Informationsverarbeitung
	Wahrgenommene hohe Vertrautheit
	Wahrgenommene geringe explizite Verbalisierbarkeit der kognitiven Informationsverarbeitung
	Wahrgenommener geringer Lernaufwand
	Wahrgenommene geringe Fehlerrate
	Wahrgenommenes gefühlsgeleitetes Entscheiden
	Wahrgenommenes magisches Erleben

Da sich laut Default-Interventionist-Theorien intuitives Handeln objektiv anhand der mentalen Beanspruchung (thematisiert von allen Forschergruppen, siehe Teilabschnitt 2.1.4) erkennen lässt (siehe Evans & Stanovich, 2013; Stanovich et al., 2014), sind die hohe zeitliche Effizienz (thematisiert von QUT- und INTUI-Forschergruppe, siehe Teilabschnitt 2.1.4) und die geringe explizite Verbalisierbarkeit während der kognitiven Informationsverarbeitung (thematisiert von QUT- und INTUI-Forschergruppe, siehe Teilabschnitt 2.1.4), sowie die Thematisierung des Unbewusstseins bei der kognitiven Informationsverarbeitung (thematisiert von allen Forschergruppen) lediglich als Korrelate intuitiver Benutzung an-

zusehen (siehe Abbildung 2.4). Diese sind nämlich bereits durch den nicht vorhandenen Bedarf an kognitivem Abkoppeln (d.h. keine Inhibition und daher keine Beanspruchung des Arbeitsgedächtnisses) impliziert, stehen aber nicht alle notwendigerweise in perfekter Korrelation zur mentalen Beanspruchung und müssen entsprechend zusammen in gleicher Ausprägung auftreten (siehe Evans & Stanovich, 2013; Stanovich et al., 2014). Durch eine derartige Reduktion können Unsauberkeiten bei der Charakterisierung intuitiver Benutzung im Zuge einer Messdefinition vermieden werden, wie sie beispielsweise durch die Verwendung der unbewussten Anwendung von Vorwissen (siehe Teilabschnitt 2.1.1), der geringen expliziten Verbalisierbarkeit oder der hohen zeitlichen Effizienz der kognitiven Informationsverarbeitung (siehe Teilabschnitt 2.1.2) als charakterisierende Merkmale intuitiver Benutzung auftreten können.

Obwohl das Merkmal der mentalen Beanspruchung während der Nutzung als definierendes objektives Merkmal intuitiver Benutzung üblicherweise auch mit einer effektiven Nutzung unter Berücksichtigung der Handlungsregulationstheorie (siehe Frese & Zapf, 1994; Hacker & Sachse, 2013) und Default-Interventionist-Theorien (siehe Ackerman & Thompson, 2017; Evans & Stanovich, 2013; Stanovich et al., 2014) korrelieren sollte (d.h. stellt dementsprechend ein Korrelat intuitiver Benutzung dar) und sich Effektivität in Folge auch direkt objektiv anhand der mentalen Beanspruchung (Cain, 2007; Hancock & Matthews, 2019) und subjektiv als wahrgenommene Zielerreichung oder wahrgenommene Fehlerrate anhand eines umfassenden Gefühls von Flüssigkeit (Ackerman & Thompson, 2017; R. Reber et al., 2002; V. Thompson, Turner, Pennycook et al., 2013) beurteilen lässt, soll in der neuen Arbeitsdefinition nicht auf das Merkmal der Effektivität zur Beurteilung des Ausmaßes intuitiver Benutzung verzichtet werden (siehe Abbildung 2.4). In der neuen Messdefinition wird Effektivität bewusst nicht als Korrelat geführt und stattdessen als zusätzliches aus pragmatischen Gründen eingeführtes definierendes Merkmal intuitiver Benutzung berücksichtigt.

Wie bereits in Teilabschnitt 2.1.1 erläutert, wird durch die explizite Berücksichtigung der tatsächlichen Effektivität (d.h. nicht nur die subjektiv wahrgenommene Effektivität in Form wahrgenommener Zielerreichung und Fehlerrate) zusätzlich sichergestellt, dass die Handlung wirklich effektiv war. Dies ist wichtig, da es bei der unbewussten Verarbeitung mithilfe des autonomen Systems auch unter bestimmten Bedingungen zu kognitiven Verzerrungen (d.h. Biases) kommen kann. Unter Biases werden im Rahmen dieser Arbeit Situationen verstanden, in denen Menschen auf Basis ihrer eigenen Überzeugungen und Stereotypen handeln, anstatt sich in ihren Entscheidungen auf Logik und Wahrscheinlichkeiten zu verlassen (V. Thompson, Evans, & Campbell, 2013). Default-Interventionist-Theorien sind der Auffassung, dass solche Biases dann zustanden kommen, wenn der autonome Geist durch ein starkes Gefühl von Flüssigkeit den beiden oberen Ebenen signalisiert, dass Typ 1 Prozesse für die kognitive Verarbeitung des aktuellen Handlungsproblems ausreichen, obwohl für die Lösung des Problems mit hoher Wahrscheinlichkeit Typ 2 Prozesse eigentlich benötigt werden würden (Evans & Stanovich, 2013; Stanovich et al., 2014; V. Thompson, Evans, & Campbell, 2013).

Beispielsweise konnten V. Thompson, Evans und Campbell (2013) unter Verwendung der Selektionsaufgabe von Foss und Dodwell (1966) zeigen, dass die Verfügbarkeit einer Anpassungsheuristik (d.h. matching heuritic) es Handelnden gestattet, schnelle Entscheidungen zu treffen, die mit einem starken Gefühl von Flüssigkeit und geringerem Überdenken der

initialen intuitiven Handlungstendenz einhergehen. Bei der Selektionsaufgabe von Foss und Dodwell (1966) werden Personen vier Karten und eine bedingte Regel (d.h. „Wenn A dann B“) dargeboten, wobei jede Karte entweder die Behauptungen in der Voraussetzung (d.h. antecedent) und der Konsequenz (d.h. consequent) bestätigt (d.h. wahr) oder negiert (d.h. falsch). Evans und Lynch (1973) konnten im Zusammenhang mit dieser Aufgabe nachweisen, dass wenn implizite Verneinungen in eine bedingte Regel einführt werden, Menschen eine hohe Tendenz besitzen (d.h. matching bias), die in der Regel genannten Karten unabhängig vom logischen Effekt zu wählen (siehe Evans, 1998). Beispielsweise wählten Personen bei der Regel „Wenn eine Karte den Buchstaben A auf einer Seite hat, hat sie nicht die Zahl 3 auf der anderen Seite“ erstaunlicherweise trotzdem Karten mit „A“ und „3“ aus, weil es sich für sie logisch „korrekt“ anfühlte.

Durch die hohe objektive Flüssigkeit und die damit verbundene geringe Beanspruchung des Arbeitsgedächtnisses kommt es hier während des Auswahlprozesses zu einem starken positiven Affekt, einem starken Gefühl von Flüssigkeit, was dem Handelnden den Eindruck vermittelt, dass die von ihm getroffenen Schlussfolgerungen und die übereinstimmenden Karten sich „richtig“ anfühlen. Dementsprechend scheinen einfache und schnelle ausgelöste Reaktionen, wie diejenigen, die von der Anpassungsheuristik ausgelöst wurden, zu einem starken Gefühl von Flüssigkeit oder Richtigkeit zu führen, obwohl das für die Handlung bereitgestellte mentale Modell für eine intuitive Benutzung (d.h. unbewusste Verarbeitung mithilfe von Typ 1 Prozessen) eigentlich nicht richtig und differenziert genug ist. An dieser Stelle können auch noch weitere einfache logische Schlussfolgerungen (d.h. modus ponens, siehe Handley, Newstead, & Trippas, 2011) und ein Prävalenzfehler (siehe Pennycook, Trippas, Handley, & Thompson, 2014) als mögliche Ursachen genannt werden, die zu einer Verzerrung des Gefühls von Flüssigkeit führen können. Für eine vertiefte Auseinandersetzung mit kognitiven Verzerrungen in Zusammenhang mit Default-Interventionist-Theorien und dem Gefühl von Flüssigkeit bzw. Richtigkeit wird auf V. Thompson, Evans und Campbell (2013), sowie Schwarz (2015) verwiesen.

Obwohl das metakognitive Gefühl von Flüssigkeit durch eine Reihe von Biases verzerrt werden kann, soll dieses dennoch als subjektives zentrales Merkmal intuitiver Benutzung in die neue Messdefinition mit aufgenommen werden (siehe Abbildung 6.12), da auf diese Weise auch die von den drei Forschergruppen genannten subjektiven Konsequenzen intuitiver Benutzung berücksichtigt werden können. Diese lassen sich auf Basis von Default-Interventionist-Theorien allesamt zu einem Gefühl von Flüssigkeit vereinen und lassen sich damit auch als Korrelate intuitiver Benutzung (d.h. nicht alle Gefühle müssen notwendigerweise auftreten und gleichermaßen zum Gefühl von Flüssigkeit beitragen) interpretieren (siehe Ackerman & Thompson, 2017). Auf diese Weise soll es, wie gehabt, möglich sein, intuitive Benutzung nicht nur objektiv anhand von Leistungskriterien (d.h. in der neuen Arbeitsdefinition messbar anhand des definierenden objektiven Merkmals der geringen Beanspruchung des Arbeitsgedächtnisses und den damit verbundenen objektiven Korrelaten), sondern auch in Form von subjektiven Zufriedenheitsindikatoren (d.h. in der neuen Arbeitsdefinition messbar anhand des definierenden subjektiven Merkmals des starken Gefühls von Flüssigkeit und den damit verbundenen subjektiven Korrelaten) evaluieren zu können. Im Zuge dessen soll jedoch auf die Abgrenzung in Vorbedingung und Konsequenzen intuitiver Benutzung, so wie es die IUUI-Forschergruppe in ihren Arbeiten thematisiert (siehe Hurtienne, 2011; Naumann & Hurtienne, 2010), verzichtet werden, da man mit der neuen Messdefinition eine Definition bereitstellen möchte, die den allge-

meinen Konsens aller drei primären Forschergruppen widerspiegeln sollte, wo QUT- und INTUI-Forschergruppe nicht explizit eine solche Einteilung vornehmen.

Unter Berücksichtigung der Erkenntnisse dieses zweiten Kapitels soll daher die folgende Arbeitsdefinition in Form einer Messdefinition formuliert werden, die in dieser Arbeit als Grundlage für die Evaluation intuitiver Benutzung genutzt wird:

> „Intuitive Benutzung ist das Ausmaß, mit dem ein Produkt mental effizient und effektiv genutzt wird, was mit einem starken metakognitiven Gefühl von Flüssigkeit einhergeht. Sie lässt sich dabei objektiv anhand der mentalen Beanspruchung bei der kognitiven Informationsverarbeitung im Arbeitsgedächtnis und den damit verbundenen objektiven Korrelaten messen, sowie subjektiv anhand des Gefühls von Flüssigkeit und den damit verbundenen subjektiven Korrelaten“ (siehe Abbildung 2.4).

Diese Messdefinition intuitiver Benutzung lässt sich als Grundlage für die summative Evaluation intuitiver Benutzung verwenden, da sich sowohl die Beanspruchung des Arbeitsgedächtnisses als auch das Gefühl von Flüssigkeit und deren beider Korrelate quantitativ beurteilen lassen. Auf diese Weise lässt sich eine zusammenfassende Bewertung des Ausmaßes intuitiver Benutzung des zu evaluierenden, vom Nutzer verwendeten Systems vornehmen. Darüber hinaus lassen sich auch entsprechend alle im Forschungsbereich zu intuitiver Benutzung vorgestellten objektiven und subjektiven Methoden unter diesen beiden definierenden Merkmalen intuitiver Benutzung einordnen (siehe Kapitel 3). Dadurch kann eine eindeutige Verknüpfung der einzelnen Evaluationsmethoden und der damit gemessenen Merkmale hergestellt werden. Bei der formativen Evaluation intuitiver Benutzung, die sich mit der Aufdeckung intuitive Benutzung beeinträchtigender Nutzungsprobleme beschäftigt, kann die neue Arbeitsdefinition außerdem eine gute Grundlage sein. Denn die kritischen Momente, in denen ein Bedarf an kognitivem Abkoppeln besteht und sich dieser in einer erhöhten Beanspruchung des Arbeitsgedächtnisses, sowie einem schwachen Gefühl von Flüssigkeit zeigt, können dem Evaluator wichtige Hinweise auf Nutzungsprobleme geben. Dieser kann sich infolgedessen an diesen Stellen vom Nutzer entsprechend qualitatives Feedback einholen.

Trotz der genannten Vorteile, die die vorgestellte Messdefinition für die Evaluation intuitiver Benutzung bringen mag, soll an dieser Stelle noch erwähnt werden, dass die anderen in diesem Kapitel vorgestellten HCI-Definitionen auch klare Vorteile besitzen. Die starke Thematisierung des Merkmals der unbewussten Anwendung von Vorwissen, wie es zentral in den Definitionen der IUUI-Forschergruppe (siehe Hurtienne, 2011) und der QUT-Forschergruppe (siehe Blackler, 2006) zu finden ist, ist beispielsweise für die Gestaltung intuitiver Benutzung entscheidend. Für die Gestaltung wird nämlich mit dem Vorwissensaspekt ein guter Ansatzpunkt bereitgestellt. Auf Basis dieses Aspekts können sich Gestalter dann überlegen, welches „universale“ Vorwissen vom späteren Nutzer des Systems mit hoher Wahrscheinlichkeit überwiegend unbewusst angewendet werden kann, um auf diese Weise in Konsequenz eine intuitive Benutzung unterstützen zu können. Für diese Überlegung kann das vorgestellte Kontinuum der Wissensquellen interessant sein (siehe Teilabschnitt 2.1.1). Im Rahmen dieser Arbeit wurden bereits Image Schemas vorgestellt (siehe Teilabschnitt 2.1.1), die als unbewusst repräsentierte Vorerfahrungen mit der Umwelt auch direkt zur Gestaltung intuitiver Benutzung von Benutzerschnittstellen

eingesetzt werden können (Hurtienne, 2011, 2017; Hurtienne, Klöckner, Diefenbach, Nass, & Maier, 2015; Hurtienne, Weber, & Blessing, 2008; Löffler, Hess, Maier, Hurtienne, & Schmitt, 2013). Einen Überblick über die Gestaltung intuitiver Benutzung mit Image Schemas liefert Hurtienne (2017). Für einen generellen Überblick über die Gestaltung intuitiver Benutzung wird auf Blackler (2018) verwiesen.

Auch die Fokussierung auf das Erleben einer intuitiven Benutzung, so wie es von der INTUI-Forschergruppe in ihren Arbeiten praktiziert wird (siehe Diefenbach & Ullrich, 2015; Ullrich, 2013, 2014), bietet mithilfe des dort vorgestellten Konzepts der INTUI-Patterns (d.h. relative Ausprägung intuitiver Benutzung in verschiedenen Merkmalen und verschiedenen Ausmaßen kann für eine differenzierte Beschreibung von Produkten genutzt werden als es beispielsweise ausschließlich anhand der mit diesem Produkt verbunden mentalen Beanspruchung möglich wäre) interessante Möglichkeiten zur Skizzierung des intendierten Nutzungserlebnisses im Design, sowie der Kategorisierung der erlebten intuitiven Benutzung mit verschiedenen Produkten (Ullrich, 2014). Anhand dieser kurzen Betrachtung sollte verdeutlicht werden, dass die im Rahmen der neuen Arbeitsdefinition vorgenommene Reduktion der, von den HCI-Forschergruppen über die Jahre identifizierten, charakterisierenden Merkmale und die Identifikation intuitive Benutzung definierender Merkmale für eine Messdefinition sinnvoll sind. Auf diese Weise kann gewährleistet werden, dass ein Evaluator immer zumindest die geringe Beanspruchung des Arbeitsgedächtnisses als definierendes objektives Merkmal bei seiner Evaluation berücksichtigt, und so das Ausmaß intuitiver Benutzung zuverlässig auf einem Kontinuum erfassen kann. Trotzdem können bei anderen Einsatzzwecken (z.B. Gestaltung intuitiver Benutzung) andere Merkmale intuitiver Benutzung an Wichtigkeit gewinnen und für diverse Verwendungszwecke entscheidende Ansatzpunkte bereitstellen.

Aufgrund der Zielsetzung dieser Arbeit eine Evaluationsmethode zu entwickeln mit der User Interfaces (speziell 3D-CUIs im Rahmen des Projekts 3D-GUIde) möglichst mit hoher zeitlicher Anwendungseffizienz formativ und summativ evaluiert werden können, kann die in diesem Kapitel formulierte Arbeitsdefinition diese Zielsetzung als Messdefinition durch die klare Abgrenzung von definierenden Merkmalen und Korrelaten intuitiver Benutzung unterstützen. Sie erlaubt es schließlich anhand der geringen Beanspruchung des Arbeitsgedächtnisses (inkl. damit verbundener Korrelate) ein objektives Evaluationskriterium, mit dem Gefühl von Flüssigkeit (inkl. damit verbundener Korrelate) ein subjektives Evaluationskriterium und mit der Effektivität ein pragmatisches Evaluationskriterium für eine zuverlässige Evaluation intuitiver Benutzung mit potentiell hoher wissenschaftlicher Güte bereitzustellen.

2.3 Zusammenfassung

In diesem zweiten Kapitel wurde eine Arbeitsdefinition in Form einer Messdefinition für intuitive Benutzung bereitgestellt. Zu diesem Zweck wurden zunächst verschiedene Definitionen intuitiver Benutzung aus dem HCI-Bereich vorgestellt und die darin enthaltenen objektiv und subjektiv beurteilbaren Merkmale intuitiver Benutzung aufgeführt (d.h. Korrelate intuitiver Benutzung). Im Zuge dessen wurde der Ursprung dieser Merkmale erläutert und intuitive Benutzung aus einer metatheoretischen Handlungsperspektive betrachtet, da es sich bei diesen Merkmalen nicht um dichotome Eigenschaften handelt,

sondern um Variablen, die während einer tatsächlichen Handlung auf einem Kontinuum variieren und so das Ausmaß intuitiver Benutzung beschreiben können.

Im Anschluss wurden verschiedene Arten von Zwei-Prozess-Theorien vorgestellt und die Einschränkungen von präemptiven und parallel-kompetitiven Theorien erläutert. Die Klasse der Default-Interventionist-Theorien wurde hierbei als aktueller Forschungsstand identifiziert und die zuvor gesammelten Korrelate intuitiver Benutzung mithilfe des damit verbundenen Drei-Instanzen-Modell des Geistes analysiert. Laut diesem Modell gibt bei der Handlungsregulation zunächst immer der autonome Geist, wie hier unbewusste Typ 1 Prozesse bezeichnet werden, eine initiale präattentive Handlungstendenz ab. Diese Handlungstendenz enthält nicht nur einen potentiell geeigneten Langzeitgedächtnisinhalt selbst, sondern auch eine präattentive metakognitive Komponente, die als Gefühl von Flüssigkeit bezeichnet wird. Diese Komponente entsteht aus einem präattentiven affektiven Richtimpuls während der unbewussten kognitiven Informationsverarbeitung (d.h. unbewusste Assimilation durch Typ 1 Prozesse) aufgrund der dabei stattfindenden objektiven Abrufflüssigkeit (d.h. Abrufschnelligkeit).

Anhand dieses Gefühls können präattentive metakognitive Überwachungsprozesse, die im Modell zusammenfassend als reflektierender Geist bezeichnet werden, entscheiden, ob die vom autonomen Geist geleistete kognitive Informationsverarbeitung zur Lösung des der Handlung zugrunde liegenden Problems zufriedenstellend ist. Bei einem starken Gefühl muss der reflektierende Geist die initiale Antworttendenz des autonomen Geistes nicht durch ein kognitives Abkoppeln (d.h. bewusste Akkommodation durch Typ 2 Prozesse) überschreiben. Bei einem schwachen Gefühl muss der reflektierende Geist hingegen ein kognitives Abkoppeln initiieren, was letztendlich dann von einem algorithmischen Geist ausgeführt wird. Für das kognitive Abkoppeln werden die in diesem Kapitel vorgestellten exekutiven Grundfunktionen des Arbeitsgedächtnis (d.h. Updating, Shifting, Inhibition) benötigt, weswegen sich dessen Größe in der mentalen Beanspruchung des Handelnden widerspiegelt. Denn je schwächer das Gefühl von Flüssigkeit und somit der Bedarf an kognitivem Abkoppeln ist, umso mehr bewusste Akkommodationsprozesse müssen für die Lösung des Handlungsproblems aufgewendet werden, was sich gleichermaßen in einer erhöhten mentalen Beanspruchung des Handelnden erkennbar macht.

Laut dem Drei-Instanzen-Modell des Geistes kann zwischen intuitiver und reflektierender kognitiver Informationsverarbeitung anhand des Bedarfs kognitiven Abkoppelns und des damit einhergehenden Gefühls von Flüssigkeit differenziert werden. Diese Einsicht wurde im Anschluss auf die zuvor gesammelten korrelierenden Merkmale intuitiver Benutzung übertragen, um diese neu unter diesen beiden zentralen definierenden Merkmalen zu gruppieren. Das Kapitel endete damit, dass basierend auf den gebildeten Gruppen eine über alle vorgestellten HCI-Definitionen intuitiver Benutzung verallgemeinernde Arbeitsdefinition gebildet werden konnte, die intuitive Benutzung als das Ausmaß versteht, mit dem ein Produkt effektiv und mental effizient genutzt wird, was mit einem starken metakognitiven Gefühl von Flüssigkeit einhergeht. Auf Grundlage dieser Messdefinition kann intuitive Benutzung im weiteren Verlauf dieser Arbeit sowohl formativ als auch summativ evaluiert werden, was Gegenstand des folgenden dritten Kapitels dieser Arbeit ist.

3 Evaluation intuitiver Benutzung

Nachdem im letzten Kapitel eine Arbeitsdefinition für intuitive Benutzung in Form einer Messdefinition bereitgestellt wurde, soll in diesem Kapitel die Frage geklärt werden, wie die intuitive Benutzung eines Systems auf Basis dieser Definition evaluiert werden kann. In diesem Kapitel wird hierzu zunächst der Begriff der Evaluation eingeführt, sowie die formative und summative Evaluation als die beiden Hauptarten der Evaluation vorgestellt. Im Anschluss wird aufgezeigt, wie diese beiden Hauptarten der Evaluation selbst zu evaluieren sind (d.h. Meta-Evaluation) und hierfür geeignete Gütekriterien (d.h. formale und nicht formale) präsentiert. Abschließend wird der aktuelle Forschungsstand bezüglich formativer und summativer Evaluationsmethoden im Forschungsbereich zu intuitiver Benutzung präsentiert und aus Anwendungsprojektsicht bezüglich seiner wissenschaftlichen Güte (d.h. Bewertung bezüglich formaler wissenschaftlicher Gütekriterien) und zeitlichen Anwendungseffizienz bewertet, sowie dabei Limitationen gesammelt, die die im Zentrum dieser Arbeit stehende Entwicklung der neuen Methode *IntuiBeat* (d.h. summative Evaluation intuitiver Benutzung mit IntuiBeat-S und formative Evaluation intuitiver Benutzung mit IntuiBeat-F) rechtfertigen. Im Zuge der Bewertung des aktuellen Forschungsstands werden geeignete formative und summative Evaluationsmethoden als Außenkriterien für die Meta-Evaluation von IntuiBeat (d.h. IntuiBeat-S und IntuiBeat-F) diskutiert.

3.1 Menschzentrierte Gestaltung interaktiver Systeme

Wenn man ein interaktives System oder, konkret im Anwendungsprojekt 3D-GUIde, eine 3D-CUI-Interaktionslösung gebrauchstauglich oder im Speziellen intuitiv benutzbar gestalten möchte, wird dafür ein bestimmtes systematisches Vorgehen benötigt. In der Norm ISO-9241 (ISO, 2011) wird dieser Prozess auch als *menschzentrierte Gestaltung interaktiver Systeme* bezeichnet, während man in der Literatur auch öfter auf den Begriff *Usability Engineering* (siehe Nielsen, 1994; Sarodnick & Brau, 2006) stößt. Aufgrund der Tatsache, dass sich intuitive Benutzung als Subkonzept von Usability versteht (siehe Teilabschnitt 2.1.1), ist dieser Prozess auch für die Gestaltung intuitiv benutzbarer Systeme anwendbar. Als paralleler Teilprozess der Entwicklung und Gestaltung eines interaktiven Systems, ergänzt er das klassische Software Engineering um die Benutzerperspektive und kann dadurch bei der Entwicklung die Umsetzung intuitiv benutzbarer Lösungen sicherstellen (Löffler et al., 2013; Wegerich, Löffler, & Maier, 2012). Die zugrunde liegenden Aspekte und Aktivitäten, sowie Empfehlungen und Anforderungen für Gestaltungsprinzipien im Rahmen eines solchen Erstellungsprozesses sind in der DIN EN ISO 9241-210 (ISO, 2011) integriert. Der darin enthaltene *Prozess zur Gestaltung gebrauchstauglicher Systeme* (siehe Abbildung 3.1) stellt bewusst den Menschen in den Mittelpunkt des Entwicklungsprozesses.

Der in der Norm beschriebene Prozess sieht vier Aktivitäten vor, welche mehrmals iterativ durchlaufen werden und durch eine Planung sowie einen Abschluss des Vorgehens

ergänzt werden. Als Iteration wird dabei eine Tätigkeit bezeichnet, bei der eine bestimmte Schrittfolge wiederholt wird, bis anfängliche Unsicherheiten beseitigt sind, ein gewünschtes Ergebnis vorliegt und ein umfassendes Verständnis über das zu entwickelnde System aufgebaut ist (König, 2012). Ein Übergang von einer Aktivität zur nächsten erfolgt erst dann, wenn die Ziele der aktuellen Aktivität erreicht wurden. Es handelt sich beim Prozess zur Gestaltung gebrauchstauglicher Systeme um einen linearen iterativen Prozess, welcher durch bedingte Querverweise nach der Aktivität *Gestaltungslösungen evaluieren* ergänzt wurde, um eine hohe Flexibilität zu gewährleisten. Des Weiteren können Überschneidungen und parallele Bearbeitung der einzelnen Aktivitäten erfolgen (König, 2012).

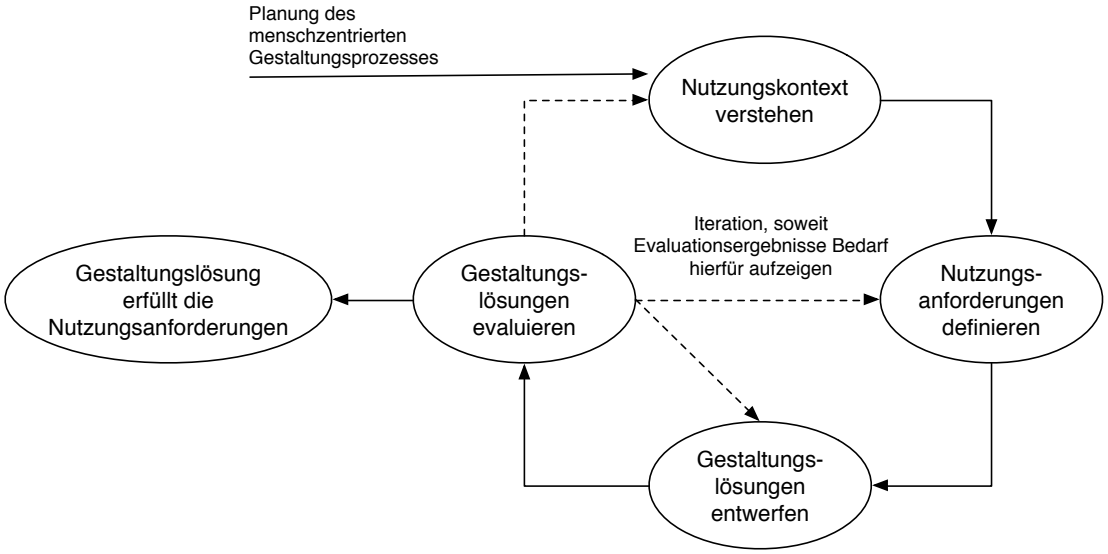


Abbildung 3.1. Prozess zur Gestaltung gebrauchstauglicher Systeme in Anlehnung an DIN EN ISO 9241-210 (ISO, 2011).

Entwickelte Konzepte, Patterns und Prototypen werden in der letzten Aktivität bezüglich ihrer Gebrauchstauglichkeit oder, wie konkret im Projekt 3D-GUIde, bezüglich ihrer intuitiven Benutzung und dem damit zusammenhängenden Erfüllungsgrad der Nutzungsanforderungen evaluiert. Die Norm ISO 9241-210 (ISO, 2011) weist hierzu an verschiedenen Stellen darauf hin, dass die Evaluation nicht als eine dedizierte Phase innerhalb des Prozesses zu sehen ist, sondern vielmehr als eine begleitende Aktivität während eines gesamten Produktlebenszyklus gesehen werden soll (Sarodnick & Brau, 2006). Eine Evaluation ist in sämtlichen Phasen des vorgestellten menschzentrierten Gestaltungsprozesses sinnvoll, denn je länger eine Evaluation aufgeschoben wird, umso größer können die Kosten infolge von Änderungen sein (Sarodnick & Brau, 2006). Die Evaluation kann dabei analytisch (d.h. Inspektion) durch Experten oder empirisch durch die direkte Einbeziehung von tatsächlichen Nutzern (d.h. Nutzertest) erfolgen (Döring & Bortz, 2016; Hegner, 2003; Sarodnick & Brau, 2006). Beim Nutzertest wird das System mit repräsentativen Endanwendern anhand realistischer Aufgaben evaluiert. Dieser stellt daher eine quasi-experimentelle Vorgehensweise mit oder ohne eingeschränkte Variation der Bedingungen dar (Wandke, 2004). Inspektionsmethoden ermöglichen eine Bewertung des Systems, indem Experten dessen Zustand ohne direkte Nutzereinbindung mithilfe eines Leitfadens oder Modells (z.B. Heuristiken und Designprinzipien) analysieren (Nielsen, 1994).

Inspektionsmethoden lassen sich aber nur schwer zur Evaluation intuitiver Benutzung einsetzen, da hierfür eine Einschätzung der Richtigkeit und Differenziertheit des mentalen Modells der Nutzer durch Experten zu erfolgen hat, um inhaltlich valide Aussagen bezüglich intuitiver Benutzung tätigen zu können. Wissen lässt sich zwar beispielsweise mithilfe kognitiver Architekturen (z.B. ACT-R, SOAR, EPIC, siehe Gray, Young, & Kirschenbaum, 1997) bis zu einem gewissen Teil abbilden und sich damit eine Inspektion intuitiver Benutzung durchführen, welche aber nur in eingeschränktem Maße (d.h. kompakte Interaktionen, begrenzte Komplexität) aussagekräftig ist, da solche Architekturen das vollständige Vorwissen eines Menschen natürlich nicht erfassen können. Für eine ausgiebige Auseinandersetzung mit kognitiver Modellierung wird auf Anderson (2013) verwiesen. Aufgrund der hohen Schwierigkeit, die intuitive Benutzung eines Systems unabhängig von direkter Nutzerbeteiligung erfassen zu können, findet man in der aktuellen Forschungsliteratur zu intuitiver Benutzung auch überwiegend empirische Evaluationsmethoden, was ein aktuelles Review von Blackler et al. (2018) zeigt. Die einzige Ausnahme bildet die von der IUUI-Forschergruppe vorgeschlagene Methode *Evalint*, die eine Checkliste bereitstellt, mit der analytisch die Berücksichtigung von Voraussetzungen für eine intuitive Mensch-Maschine-Interaktion sichergestellt werden können (Mohs, Hurtienne, Scholz, & Rotting, 2006). Wenn im Rahmen dieser Arbeit also fortan von Evaluationsmethoden gesprochen wird, sind immer empirische Methoden im Sinne eines Nutzertests gemeint. Auf Inspektionsmethoden wird im Rahmen dieser Arbeit nicht weiter eingegangen.

3.2 Evaluation in der HCI

Der Begriff *Evaluation* wird gleichermaßen für die verwendete Evaluationsmethode als auch für deren Ergebnis verwendet und bezeichnet allgemein eine systematische und gleichzeitig möglichst objektive Bewertung eines geplanten, laufenden oder abgeschlossenen Projekts im Rahmen des menschzentrierten Gestaltungsprozesses (Sarodnick & Brau, 2006). Anders ausgedrückt, versteht man darunter das systematische Sammeln, Auswerten und Interpretieren von Daten, um eine objektive, reliable und valide Bewertung des Evaluationsgegenstandes zu ermöglichen (Görner & Ilg, 1993; Hegner, 2003). Die Deutsche Gesellschaft für Evaluation (DeGEval) wird hierbei noch etwas konkreter und definiert Evaluation als die *systematische Untersuchung der Güte oder des Nutzens eines Evaluationsgegenstands*, was auch als Arbeitsdefinition im Rahmen dieser Arbeit fungieren soll. Sie charakterisiert Evaluation anhand der folgenden drei Eigenschaften (DeGEval, 2016):

- Nachvollziehbares systematisches Vorgehen (d.h. auf bestimmtes Evaluationsziel ausgerichtet) auf Basis von empirisch gewonnenen Daten.
- Transparente, kriteriengeleitete Bewertung eines Evaluationsgegenstandes, die entweder vor dem Hintergrund (1) eines *bestimmten Verwendungskontextes* oder *Nutzens* (z.B. Feststellung, ob eine Interaktionslösung stärker intuitiv benutzbar als eine andere ist oder das Ableiten von Verbesserungsvorschlägen für eine Interaktionslösung) oder (2) *übergreifend zur Sicherstellung der Güte* vorgenommen wird.
- Anwendbarkeit auf verschiedene Evaluationsgegenstände. Hier sind insbesondere Projekte, Systeme, Produkte, Maßnahmen, Programme, Organisationen und Evaluationen selbst zu nennen.

3.2.1 Durchführung einer Evaluation

Eine jede Evaluation findet üblicherweise kriteriengeleitet in einem vierstufigen Prozess statt, um einen möglichst transparenten Bewertungsprozess zu ermöglichen. Die vier Stufen dieses formalen Prozesses sind die Folgenden (siehe Baumgartner, 1999; DeGEval, 2016; Hegner, 2003; Scriven, 1967, 1991):

1. **Formulierung von Evaluationskriterien** Evaluationskriterien sind jene Merkmale des Evaluationsgegenstands, in denen sich dessen Nutzen oder dessen Güte zeigen. Die Definition, Auswahl, Priorisierung und Dokumentation stellen somit die essentielle Voraussetzung für eine systematische Bewertung der Evaluation dar (DeGEval, 2016).
2. **Formulierung von Außenkriterien** Für die im ersten Schritt eingeführten Evaluationskriterien werden Außenkriterien definiert, die vorgeben, welche Ausprägung ein Evaluationskriterium (d.h. Zielwert) aufweisen sollte, um als erfüllt zu gelten (DeGEval, 2016). Außenkriterien werden unabhängig vom Evaluationsgegenstand erhoben (d.h. befinden sich außerhalb der Testsituation) und müssen in irgendeiner direkten oder indirekten Weise das Evaluationskriterium, dessen gewünschte Ausprägung sie bestimmen, repräsentieren oder widerspiegeln (Lienert & Raatz, 1998).
3. **Messung und Vergleich (d.h. Analyse)** Der Evaluationsgegenstand wird unter der Anwendung der Evaluationskriterien untersucht, gemessen und mit den formulierten Außenkriterien verglichen (DeGEval, 2016).
4. **Werturteil (d.h. Synthese)** Abschließend werden die einzelnen Ergebnisse zu einem einheitlichen Werturteil der Evaluation (z.B. gut, ausreichend) verknüpft (DeGEval, 2016).

3.2.2 Ziele der Evaluation

Neben der systematischen, kriteriengeleiteten Bewertung des Evaluationsgegenstands hat eine Evaluation immer auch eine *bestimmte Verwendungsabsicht oder einen speziellen Leistungsschwerpunkt*, auf die oder den sie in Planung und Durchführung ausgerichtet werden muss. Zur Abgrenzung der Ebene des evaluierten Gegenstands, spricht man hier von *Evaluationszielen* (d.h. Evaluationsfunktionen, Evaluationszwecke, siehe Döring & Bortz, 2016) anstelle von Evaluationskriterien (DeGEval, 2016). Laut Döring und Bortz (2016) existieren in der Literatur eine Reihe von Auffassungen, welche Evaluationsziele das genau sein können (DeGEval, 2016; Stockmann, 2000; Widmer & De Rocchi, 2012). Grob lässt sich hier zwischen den folgenden drei Zielen unterscheiden, die sich in den Aufzählungen unterschiedlicher Autoren wiederfinden (z.B. Auf der Heide, 1993; Gediga & Hamborg, 2002; Hegner, 2003; Sarodnick & Brau, 2006):

Vergleichende Evaluation („Which is better?“) Bei dieser Fragestellung werden mindestens zwei unterschiedliche Evaluationsgegenstände hinsichtlich bestimmter Evaluationskriterien verglichen (Auf der Heide, 1993; Hegner, 2003; Sarodnick & Brau, 2006).

Bewertende Evaluation („How good?“) Bei dieser Fragestellung geht es darum, die Ausprägung eines bestimmten Evaluationskriteriums durch den Vergleich mit einem geeigneten Außenkriterium zu prüfen (Auf der Heide, 1993; Hegner, 2003; Sarodnick & Brau, 2006).

Analysierende Evaluation („Why bad?“) Bei dieser Fragestellung geht es darum, Hinweise auf Schwachstellen zu erhalten, um direkte Verbesserungsvorschläge des evaluierten Gegenstands liefern zu können (Auf der Heide, 1993; Hegner, 2003; Sarodnick & Brau, 2006).

Hinsichtlich dieser drei Evaluationsziele lässt sich bei der Evaluation von Evaluationsgegenständen grundsätzlich zwischen *gestaltend-formativer* (d.h. induktive Evaluation) und *bilanzierend-summativer* Evaluation (d.h. deduktive Evaluation) unterscheiden (DeGEval, 2016; Döring & Bortz, 2016; Dumas, Dumas, & Redish, 1999; Hegner, 2003; Sarodnick & Brau, 2006; Scriven, 1967). Beide Hauptarten der Evaluation werden nun nacheinander vorgestellt und ihre Besonderheiten aufgezeigt.

3.2.3 Gestaltend-formative Evaluation

Bei dieser Hauptart der Evaluation steht das dritte Evaluationsziel (d.h. analysierende Evaluation) im Vordergrund (Hegner, 2003) und sie besitzt daher hauptsächlich eine Optimierungsfunktion (Döring & Bortz, 2016). Die gestaltend-formative Evaluation begleitet dazu die Gestaltung des Evaluationsgegenstands kontinuierlich und hilft dabei den Stakeholdern den Evaluationsgegenstand vor dem Hintergrund eines bestimmten Verwendungskontextes (z.B. Gestaltung intuitiver Benutzung) zu verbessern und die hierfür benötigten Ressourcen möglichst gut einzusetzen (DeGEval, 2016). Im Rahmen des iterativen menschzentrierten Gestaltungsprozesses nach DIN EN ISO 9241-210 (ISO, 2011) findet diese Art der Evaluation deswegen während der eigentlichen Entwicklung des Systems statt. Evaluationsergebnisse können auf diese Weise direkt und kontinuierlich in die Weiterentwicklung des Systems einfließen. Dadurch können Prototypen bzw. entwickelte Vorabversionen analysiert, frühzeitig Probleme aufgedeckt und darauf basierend Verbesserungsmöglichkeiten abgeleitet werden (Döring & Bortz, 2016; Sarodnick & Brau, 2006). Die formative Evaluation bedient sich daher überwiegend qualitativer Daten (z.B. Problembeschreibungen) und hilft die Systemnutzung eines sich in der Entwicklung befindenden Systems durch Zwischenergebnisse laufend zu verbessern (Döring & Bortz, 2016; Hartson, Andre, & Williges, 2001).

Um entsprechende Verbesserungsmöglichkeiten ableiten zu können und im Falle des Projekts 3D-GUIde Interaktionslösungen solange optimieren zu können, bis diese als Interaktionspatterns festgeschrieben werden können, müssen Probleme bei der Systemnutzung hinsichtlich ihrer Ursache analysiert und klassifiziert werden. Dies ist wichtig, da generell nicht davon auszugehen ist, dass alle gefundenen Probleme die gleichen Maßnahmen erfordern und es sich wirklich um ein echtes Problem handelt (Frese & Zapf, 1994; Zandbergen, 2015; Zapf et al., 1989). Bei einer solchen Klassifikation hilft es die Ursachenanalyse zu standardisieren und dadurch einer zu flachen Ursacheninterpretation vorzubeugen (Hamborg, Hoemske, & Ollermann, 2006). Des Weiteren kann anhand einer solchen Analyse auch entschieden werden, ob es sich bei den gefundenen Problemen um *reale Nutzungsprobleme bezüglich intuitiver Benutzung* handelt (d.h. beeinträchtigen wirklich die intuitive

Benutzung eines Systems und nicht nur die physische oder zeitliche Effizienz bei der Handlungsausführung, Merkmale, die zwar mit Usability in Verbindung stehen aber nicht mit intuitiver Benutzung, siehe Teilabschnitt 2.1.1) oder lediglich um *Ineffizienzen* (d.h. es tritt bei der Systemnutzung kein Problem auf, sondern die Systemnutzung ist lediglich zeitlich ineffizient) (Zapf et al., 1989).

Aus einer handlungsorientierten Perspektive sind bei einer formativen Evaluation reale gefundene Probleme immer dadurch gekennzeichnet, dass ein angestrebtes Ziel nicht erreicht wurde (Zapf et al., 1989). Ein System hat keine Ziele und kann lediglich die von ihm erwartete Aufgabe nicht erfüllen. Man spricht an dieser Stelle von einem *Funktionsproblem*, wenn ein System seinen Zweck im Rahmen einer bestimmten Aufgabe nicht erfüllen kann. Die Funktionalität eines Systems bezieht sich hierbei auf die Beziehung von Aufgabe und System (siehe Abbildung 3.2). Funktionsprobleme können sich unterschiedlich auf den Handlungsprozess auswirken. Sie können im Extremfall dazu führen, dass der Nutzer das Durchführen seiner Aufgaben oder allgemein die Systemnutzung abbricht oder er zumindest dazu gezwungen ist, sich mit Kompensationsstrategien (d.h. Workarounds) zu behelfen, die zu Fehlern in der Systemnutzung führen. Sie können daher als Diskrepanz (d.h. Mismatch) zwischen Aufgabe und System begriffen werden (Frese & Zapf, 1994; Zandbergen, 2015; Zapf et al., 1989).

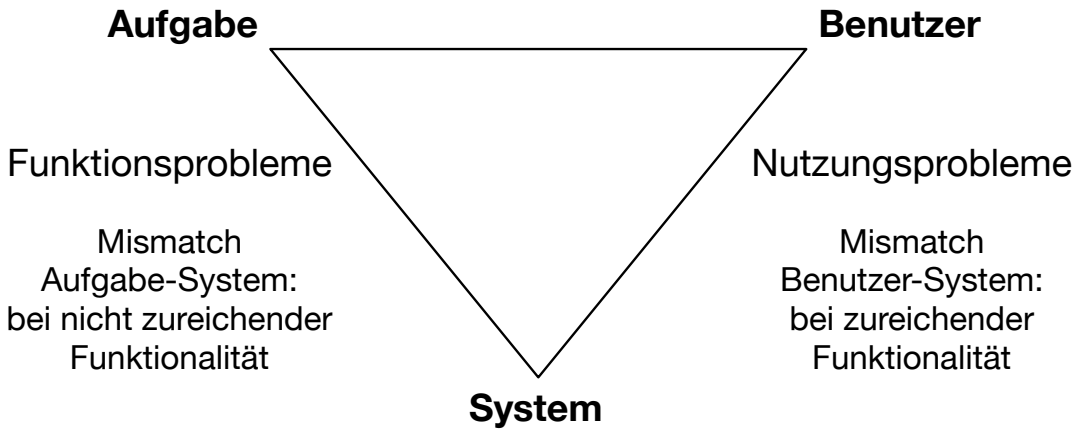


Abbildung 3.2. Funktionsprobleme und Nutzungsprobleme im Kontext von Aufgabe, Benutzer und System in Anlehnung an Zapf, Brodbeck und Prümper (1989).

Im Sinne eines menschzentrierten Gestaltungsprozesses DIN EN ISO 9241-210 (ISO, 2011) ist die Gewährleistung eines funktionstüchtigen Systems jedoch nicht ausreichend (Sarodnick & Brau, 2006). Aus einer menschzentrierten Perspektive ist vielmehr die Passung zwischen Benutzer und System entscheidend, da auch ein funktionierendes System für den Benutzer zu Problemen führen kann, wenn dieser beispielsweise nicht das entsprechende mentale Modell für die Arbeitsaufgabe mitbringt. Diese Diskrepanz zwischen System und Benutzer wird auch als *Nutzungsproblem* bezeichnet (siehe Abbildung 3.2) und soll ausdrücken, dass es sich um ein Problem des Gesamtsystems handelt. Die Ursache des Problems kann in vielen Fällen nämlich keinem der beiden Teilsysteme (d.h. Benutzer oder System) eindeutig zugeordnet werden (Zapf et al., 1989). Aufgrund der Tatsache,

dass Nutzungsprobleme erst durch die eigentliche Systeminteraktion eines Nutzers sichtbar werden, stehen diese auch im Mittelpunkt einer formativen Evaluation im Rahmen des menschzentrierten Gestaltungsprozesses. Die Durchführung einer Ursachenanalyse bei Funktionsproblemen (z.B. Softwarebugs), die unabhängig von einem bestimmten Nutzer existieren, liegt daher überwiegend in der Zuständigkeit des klassischen Software Engineerings. Für einen Überblick über verschiedene Möglichkeiten der Ursachenanalyse aus der Perspektive des Software Engineerings wird auf Myers, Sandler und Badgett (2011) verwiesen.

Bei Nutzungsproblemen muss für eine tiefgreifende Ursachenanalyse festgestellt werden, was den Nutzer aus psychologischer Perspektive daran hindert, sein Handlungsziel trotz eines technisch funktionierenden Systems zu erreichen. Erst durch das Beheben dieser *realen Nutzungsprobleme* kann eine gebrauchstaugliche (d.h. Nutzungsproblem beeinträchtigt eine effektive, effiziente und zufriedenstellende Systemnutzung) oder eine intuitive Benutzung des Systems (d.h. Nutzungsproblem beeinträchtigt eine effektive, mental effiziente und zufriedenstellende Systemnutzung) gewährleistet werden. Im Folgenden ist daher immer Nutzungsproblem gemeint, wenn von einem Fehler oder einem Problem gesprochen wird. In der Literatur werden für die Ursachenanalyse eine Reihe von Fehlertaxonomien angeboten. Jedoch konzentrieren sich die meisten dieser Taxonomien (z.B. Hamborg et al., 2006; Hartson, Andre, Williges, & Van Rens, 1999; D. A. Norman & Shallice, 1986) lediglich auf die beobachtbaren Ursachen von Fehlern und stützen ihre Klassifikation auf Aufgabencharakteristika, sowie Mängel im System. Die Ursache für Fehler liegt gemäß dieser Taxonomien somit auf Systemseite und Fehler können einfach durch eine entsprechende Neugestaltung des Systems behoben werden (Sarnikar & Murphy, 2012). An dieser Stelle soll noch explizit darauf hingewiesen werden, dass mithilfe solcher Taxonomien die Ursache von Nutzungsproblemen klassifiziert und festgehalten, aber nicht deren Konsequenz beurteilt wird. Beispielsweise kann ein Nutzungsproblem eine unbewusste Ursache habe, aber aufgrund der Tatsache, dass der Nutzer das Problem erkennt, eine bewusste Korrekturhandlung nach sich ziehen.

Eine Ausnahme bildet die *handlungsorientierte Fehlertaxonomie* von Zapf et al. (1989). Diese postuliert aus einer psychologischen Perspektive, dass Fehler nicht durch das System oder den Benutzer alleine verursacht werden und damit nicht unabhängig vom Nutzer analysierbar sind. Die eigentlichen Fehler (d.h. Nutzungsprobleme) kommen erst durch eine Interaktion von Mensch und System zustande und deren Ursache kann partiell dem kognitiven Apparat des Menschen zugeordnet werden. Diese Taxonomie wurde bei formativen Evaluationsmethoden im Rahmen eines menschzentrierten Gestaltungsprozesses bereits mehrfach eingesetzt (z.B. Barendregt, Bekker, Bouwhuis, & Baauw, 2006; Fu, Salvendy, & Turley, 2002; Heimbeck, Frese, Sonnentag, & Keith, 2003; Zandbergen, 2015; Zapf, Brodbeck, Frese, Peters, & Prümper, 1992), da sie besonders ausführlich beschriebene Fehlerkategorien bietet und dadurch eine besonders umfangreiche Ursachenanalyse ermöglicht (Zandbergen, 2015). Des Weiteren kann mit der Taxonomie sehr leicht zwischen Funktionsproblemen und Nutzungsproblemen unterschieden werden, da nur Probleme, die sich anhand der Taxonomie klassifizieren lassen, Nutzungsprobleme aus einer menschzentrierten Gestaltungsperspektive darstellen. Nur diese Probleme fallen damit in die Zuständigkeit des Usability Engineerings. Bei den damit nicht klassifizierbaren Problemen handelt es sich um Funktionsprobleme, die daher in der Zuständigkeit des Software Engineerings liegen und an dieses weitergeleitet werden können. Im Rahmen dieser Ar-

beit wird deswegen nur diese Taxonomie genauer betrachtet und für einen Überblick über andere Fehlertaxonomien auf Sarnikar und Murphy (2012) verwiesen.

Die handlungsorientierte Fehlertaxonomie geht davon aus, dass die kognitive Informationsverarbeitung auf den bereits in Kapitel 2 beschriebenen drei Ebenen (d.h. sensomotorische Ebene, perzeptiv-begriffliche Ebene, intellektuelle Ebene) reguliert wird. Es können daher auch Nutzungsprobleme hinsichtlich dieser Ebenen und somit nach dem Ausmaß unterschieden werden, wie wahrscheinlich diese Probleme die intuitive Benutzung des Systems beeinträchtigen werden. Wie bereits in Kapitel 2 angesprochen, stellt intuitive Benutzung im Gegensatz zur Gebrauchstauglichkeit eines Systems, insbesondere die mental effiziente Benutzung des Systems als charakterisierendes Merkmal in den Mittelpunkt und lässt sich daher objektiv in seinem Ausmaß anhand der mentalen Beanspruchung des Nutzers bewerten. Die mit einer formativen Evaluationsmethode identifizierten Nutzungsprobleme sollten dementsprechend auch Aspekte beschreiben, die speziell die intuitive Benutzung des Systems beeinträchtigen und nicht nur seine generelle Gebrauchstauglichkeit. Unter einem „realen Nutzungsproblem“ werden demzufolge im Rahmen dieser Arbeit alle Aspekte verstanden, die aus einem „Mismatch“ zwischen System und Nutzer entstehen und dabei speziell eine hohe mentale Effizienz bei der Systemnutzung gefährden.

Die handlungsorientierte Fehlertaxonomie kann aufgrund ihrer psychologischen Ausrichtung bei der Bestimmung speziell für die intuitive Benutzung relevanter und damit realer Nutzungsprobleme helfen. Da nur die oberen beiden Regulationsebenen innerhalb der Taxonomie mit mentaler Beanspruchung und einem Bedarf an kognitivem Abkoppeln verbunden sind, beeinträchtigen auch nur Nutzungsprobleme, die diesen Ebenen zugeordnet werden können, die intuitive Benutzung des Systems (siehe Tabelle 3.1). Nutzungsprobleme, die mithilfe der handlungsorientierten Fehlertaxonomie der dritten Regulationsebene (d.h. sensomotorische Ebene erfordert aufgrund vollständiger Assimilation kein kognitives Abkoppeln) zugeordnet werden können, beeinträchtigen lediglich die motorische Effizienz des Systems und spiegeln sich daher lediglich in dessen Gebrauchstauglichkeit wider (siehe Kapitel 2.2). Diese Probleme können demnach nicht als reale Nutzungsprobleme im Sinne einer intuitiven Benutzung bezeichnet werden. Mithilfe dieser Klassifikation können daher formative Evaluationsergebnisse gewonnen werden, auf deren Basis die intuitive Benutzung des Systems verbessert werden kann.

Tabelle 3.1. *Handlungsorientierte Taxonomie der Nutzungsprobleme in Anlehnung an Zapf, Brodbeck und Prümper (1989).*

Regulationsgrundlage (Fehler aufgrund fehlenden Vorwissens)	Wissensfehler		
Regulationsebenen	Schritte im Handlungsprozess		
	Ziele/Planung	Gedächtnis/Monitoring	Rückmeldung
Intellektuelle Ebene (Fehler bei der Akkommodation)	Denkfehler	Merk-/Vergessensfehler	Urteilsfehler
Perzeptiv-begriffliche Ebene (Fehler aufgrund des Wechsels zwischen Assimilation und Akkommodation)	Gewohnheitsfehler	Unterlassensfehler	Erkennensfehler
Sensomotorische Ebene (Fehler bei der Assimilation)	Bewegungsfehler		

Um eine stärkere Differenzierung bei der Ursachenanalyse zu erreichen, lassen sich im Rahmen der handlungsorientierten Fehlertaxonomie Nutzungsprobleme auch nach Schritten im Handlungsprozess innerhalb dieser drei Ebenen einteilen. Ein derartig detaillierte Fehlertaxonomie hilft gezielt die Ursachen von Nutzungsproblemen zu analysieren und unter Berücksichtigung der kognitiven Informationsverarbeitung des Nutzers entsprechende Gegenmaßnahmen abzuleiten. Der Handlungsprozess wird in der Taxonomie von den ursprünglichen von Hacker (1986) festgelegten fünf sequentiellen Schritten (d.h. Zielsetzung, Orientierung, Entwerfen und Entscheiden, Ausführung, Feedbackverarbeitung), die jeder Handelnde im Rahmen der Handlungsregulation durchläuft, in drei Schritte (d.h. Zielbildung, Zielspeicherung und Feedbackverarbeitung) zusammengefasst (siehe Teilabschnitt 2.1.1).

Zapf et al. (1989) rechtfertigen diese Reduktion dadurch, dass zu Beginn einer jeden Handlung zunächst ein Ziel definiert und darauf basierend ein Handlungsplan zur Zielrealisierung entworfen werden muss. Dieser Plan muss solange im Arbeitsgedächtnis präsent gehalten werden, bis er abgerufen werden kann. Am Ende der Handlung erhält der Handelnde Feedback, inwiefern das angestrebte Ziel erreicht wurde. Nutzungsprobleme können entweder bei Planungsprozessen, Planspeicherungs- /Abrufprozesen oder Feedbackprozessen auftreten. Außerdem wird neben den drei Regulationsebenen noch die Regulationsgrundlage (d.h. mentales Modell) selbst als mögliche Fehlerklasse eingeführt, da einem Nutzer auch gänzlich das Vorwissen für die Erledigung der Arbeitsaufgabe mit dem System fehlen kann oder er sich zuvor überhaupt kein explizites Handlungsziel gesetzt hat, was überprüfbar wäre. Die aus den hierarchischen Regulationsebenen (inkl. Regulationsgrundlage) und dem sequentiellen Handlungsverlauf resultierenden acht Fehlertypen (siehe Tabelle 3.1), die zur Klassifikation im Rahmen einer Ursachenanalyse genutzt werden können, sollen nun nachfolgend kurz beschrieben und an einfachen Beispielen verdeutlicht werden.

Fehler in der Regulationsgrundlage bedeuten, dass der Nutzer kein Vorwissen für die durchzuführende Handlung besitzt und entsprechend im Zuge der Handlungsregulation überhaupt kein mentales Modell (d.h. operatives Abbildsystem) für die Handlung bereitstellen kann. Diese Fehler werden auch als **Wissensfehler** bezeichnet (Zandbergen, 2015; Zapf et al., 1992; Zapf et al., 1989). Beispiel: Ein Nutzer hat keine Ahnung, wie man mit einem CAD-Programm überhaupt arbeitet. Eine fehlerfreie Handlungsregulation wird bereits durch die fehlende Regulationsgrundlage verhindert.

Fehlern auf der intellektuellen Regulationsebene ist gemeinsam, dass sie bei der bewussten kognitiven Informationsverarbeitung (d.h. Akkommodation) einer Handlung entstehen (siehe Tabelle 3.1), da die Handlung für den Nutzer überwiegend neu ist (Zandbergen, 2015; Zapf et al., 1992; Zapf et al., 1989) und deswegen ausschließlich Typ 2 Prozesse zur Bereitstellung des mentalen Modells genutzt werden können (Stanovich et al., 2014). Aufgrund des damit verbundenen vollständigen kognitiven Abkoppelns (siehe Teilabschnitt 2.2) ist die mentale Beanspruchung auf dieser Ebene hoch und das subjektive Gefühl von Flüssigkeit bzw. Richtigkeit schwach (siehe Stanovich et al., 2014). Daher können Nutzungsprobleme, die der intellektuellen Regulationsebene zugeordnet werden können, die intuitive Benutzung des Systems beeinträchtigen. Im Zuge der formativen Evaluation intuitiver Benutzung müssen diese Probleme also erkannt und im Rahmen des menschzentrierten Gestaltungsprozesses beseitigt werden.

Denkfehler Diese Fehler treten bei der Planungsphase auf der intellektuellen Regulationsebene auf. Im Gegensatz zu Wissensfehlern besitzt der Nutzer die Regulationsgrundlage für die Handlung in seinem Langzeitgedächtnis (d.h. Vorwissen), jedoch stellt der Nutzer darauf basierend falsche oder unrealistische Handlungsziele auf, setzt Teilhandlungen falsch zusammen oder trifft falsche Entscheidungen zwischen den Teilzielen (Zandbergen, 2015; Zapf et al., 1992; Zapf et al., 1989). Aus diesem Grund führen entsprechende Handlungspläne nur durch den Einsatz großer mentaler Beanspruchung zum Ziel, da das Handlungsproblem aufgrund der mangelnden Richtigkeit und Differenziertheit des aufgebauten mentalen Modells nicht durch Assimilation lösbar ist. Stattdessen muss komplett auf bewusste Akkommodationsprozesse zurückgegriffen werden (siehe Stanovich et al., 2014), um ein vollständiges kognitives Abkoppeln leisten zu können (siehe Teilabschnitt 2.2). Beispiel: Ein Nutzer erstellt mit einem CAD-Programm einen Motor für eine unbekannte Maschine mit vier Zylinderbohrungen. Nach langem Arbeiten und investierten geistigen Ressourcen merkt der Nutzer aber, dass man mit vier Bohrungen nicht auskommt. Der Nutzer muss seinen Entwurf noch einmal von vorn anfangen, weil er sich das falsche Ziel gesetzt hat und er sein eigentliches Ziel so nicht erreichen kann.

Merk- und Vergessensfehler Merk- und Vergessensfehler finden während der Planspeicherung bzw. dem Monitoring auf der intellektuellen Regulationsebene statt. Der Nutzer konnte zwar mithilfe bewusster Akkommodationsprozesse ein für die Handlung prinzipiell angemessenes mentales Modell erstellen, vergisst jedoch vor der eigentlichen Ausführung einen Teil des Handlungsplans (Zandbergen, 2015; Zapf et al., 1992; Zapf et al., 1989). Dementsprechend muss er zur Kompensation erneut auf bewusste Akkommodationsprozesse zurückgreifen (siehe Stanovich et al., 2014), um ein vollständiges kognitives Abkoppeln leisten zu können (siehe Teilabschnitt 2.2). Beispiel: Der Nutzer druckt sich seinen Motorentwurf aus und stellt dann beim Durchsehen fest, dass eine der geplanten Zylinderöffnungen vergessen wurde.

Urteilsfehler Urteilsfehler finden während der Feedbackverarbeitung einer Handlung auf der intellektuellen Regulationsebene statt, also nachdem der Nutzer Feedback von Systemseite erhalten hat. Falls ein Nutzer dieses Feedback aufgrund seines vorliegenden mentalen Modells nicht versteht, nicht wahrnimmt oder es einfach falsch interpretiert, liegt ein Urteilsfehler vor (Zandbergen, 2015; Zapf et al., 1992; Zapf et al., 1989). Die Bezeichnung Urteilsfehler wurde gewählt, um kenntlich zu machen, dass es sich hierbei um einen bewussten Prozess handelt (Zandbergen, 2015; Zapf et al., 1992; Zapf et al., 1989), da aufgrund des neuartigen Feedbacks bewusste Akkommodationsprozesse zum Einsatz kommen (siehe Stanovich et al., 2014) und daher ein vollständiges kognitives Abkoppeln nötig wird (siehe Teilabschnitt 2.2). Beispiel: Muss ein Nutzer den Entwurf seines Motors in ein anderes Format umwandeln, kann es passieren, dass das System ihm als Rückmeldung lediglich „Fehler.06“ gibt, was er noch nie gehört hat und nicht interpretieren kann.

Fehlern auf der perzeptiv-begrifflichen Regulationsebene ist gemeinsam, dass sie bei gut beherrschten Handlungen auftreten können. Es werden keine neuartigen Pläne erstellt oder verarbeitet, sondern es wird auf dieser Zwischenebene anhand von Urteils- und Klassifikationsvorgängen entschieden (Semmer & Pfäfflin, 1978), ob unbewusste Assimilationsprozesse oder bewusste Akkommodationsprozesse für die kognitive Informationsverarbeitung

benötigt werden. Anders ausgedrückt: Das aktuell verfügbare mentale Modell muss aufgrund für die Lösung des Handlungsproblems nicht ausreichender Richtigkeit und Differenziertheit durch bewusste Akkommodationsprozesse an die jeweilige Handlungssituation angepasst werden. Die kognitive Informationsverarbeitung auf dieser Ebene ist demnach als bewusstseinsfähig zu bezeichnen, da, falls die unbewusste Assimilation scheitert (z.B. Nutzer vertippt sich), für diese Anpassung (z.B. Korrektur der Eingabe) bewusste Kontrolle (d.h. Akkommodation) notwendig ist (Zacher, 2017; Zandbergen, 2015; Zapf et al., 1992; Zapf et al., 1989). Es kommt auf dieser perzeptiv-begrifflichen Ebene zu einem unvollständigen kognitiven Abkoppeln in Abhängigkeit der Stärke des Gefühls von Flüssigkeit (siehe Teilabschnitt 2.2), weswegen die mentale Beanspruchung auf dieser Ebene variieren kann, aber auf jeden Fall größer ist, als bei vollständiger Assimilation auf der sensomotorischen Ebene und kleiner als bei vollständiger Akkommodation auf der intellektuellen Ebene (siehe Stanovich et al., 2014).

Nutzungsprobleme, die der perzeptiv-begrifflichen Ebene zugeordnet werden können, können dementsprechend die intuitive Benutzung des Systems beeinträchtigen. Im Zuge der formativen Evaluation intuitiver Benutzung müssen diese also erkannt und im Rahmen des menschzentrierten Gestaltungsprozesses beseitigt werden.

Gewohnheitsfehler Gewohnheitsfehler finden auf der perzeptiv-begrifflichen Ebene während der Handlungsplanung statt. Sie treten auf, wenn ein Nutzer eine Handlung korrekt ausführt, aber in einer falschen Situation. Anders ausgedrückt: Der Nutzer hat die nötige Regulationsgrundlage für die Handlung und muss diese zum Aufbau eines komplexitätsangemessenen mentalen Modells für die jeweilige Handlungssituation lediglich mit Umgebungsinformationen durch unbewusste Assimilationsprozesse ergänzen, was ihm nicht gelingt (Zandbergen, 2015; Zapf et al., 1992; Zapf et al., 1989). In solchen Fällen wird dann ein Standardwert einer ähnlichen Situation im Rahmen eines unvollständigen kognitiven Abkoppelns verwendet (siehe Teilabschnitt 2.2), welcher nicht passend sein kann und es dadurch zu einem Gewohnheitsfehler kommt (siehe Kopp & Mandl, 2005). Beispielsweise treten Gewohnheitsfehler bei Nutzern auf, wenn sie von einem alten CAD-Programm (z.B. Autodesk AutoCAD) zu einem anderen neuen CAD-Programm (z.B. Rhinoceros 3D) wechseln, und versuchen die gleichen Funktionstasten für die Konstruktion zu verwenden.

Unterlassensfehler Unterlassensfehler finden auf der perzeptiv-begrifflichen Ebene während der Planspeicherung bzw. dem Monitoring statt. Dieser Fehlertyp tritt auf, wenn ein Nutzer einen Teilplan, den er schon mehrmals richtig im gleichen Handlungskontext ausgeführt hat, vergisst auszuführen oder nur teilweise ausführt. Anders ausgedrückt: Die Umgebungsinformationen wurden bei der Umsetzung des unvollständigen kognitiven Abkoppelns zwar durch Assimilationsprozesse unbewusst in das mentale Modell integriert (siehe Teilabschnitt 2.2), aber beim Abrufen des Handlungsplans geht etwas schief, weswegen trotzdem bewusste Akkommodationsprozesse zur Kompensation involviert werden müssen (Frese & Zapf, 1994). Gründe für das Scheitern der Assimilation können Ablenkungen oder bereits die Tatsache sein, dass der Nutzer gedanklich schon beim nächsten Handlungsschritt ist (Zandbergen, 2015; Zapf et al., 1992; Zapf et al., 1989). Beispielsweise vergisst ein Nutzer seine Motorzeichnung zu sichern bevor er sein CAD-Programm schließt, was er normalerweise immer macht, da er gedanklich schon den Feierabend mit Freunden plant.

Erkennensfehler Erkennensfehler finden auf der perzeptiv-begrifflichen Ebene während der Feedbackphase statt. Diese Art von Fehlern tritt auf, wenn ein Nutzer das Systemfeedback nicht wahrnimmt oder versteht, obwohl er das Systemfeedback schon öfter richtig wahrgenommen und interpretiert hat. Anders ausgedrückt: Das mentale Modell reicht nicht für eine unbewusste Interpretation des Feedbacks und eine damit verbundene Wissensaktualisierung aus, weswegen der Nutzer bewusste Prozesse zur Interpretation benötigt und dementsprechend ein unvollständiges kognitives Abkoppeln stattfindet (siehe Teilabschnitt 2.2). Der Unterschied zu Urteilsfehlern besteht darin, dass es sich nicht um neuartiges Feedback handelt, sondern um Feedback, das bereits schon einmal richtig in der Handlungsregulation berücksichtigt wurde (Zandbergen, 2015; Zapf et al., 1992; Zapf et al., 1989). Beispielsweise ist einem CAD-Anwender normalerweise klar, dass sein System bei der Standardselektion nur die sichtbaren Teile (z.B. Motorhaube) eines Objekts (z.B. Motor) mit auswählt. Der Nutzer kann aber einstellen, dass stattdessen alle Teilobjekte (z.B. Schrauben) ausgewählt werden. Beim Neustart des Systems ist jedoch wieder die Standardselektion aktiv. Obwohl der Nutzer dies weiß, passiert es ihm öfters, dass er beim Verschieben des Motors nur den äußeren Teil des Motors verschiebt, die Schrauben aber auf ihrem Platz bleiben.

Fehlern auf der sensomotorischen Ebene ist gemeinsam, dass sie bei der automatische Umsetzung von motorischen Aktionsfolgen und der Assimilation von Kontextinformationen durch unbewusst aktiviertes Vorwissen auftreten (Zandbergen, 2015; Zapf et al., 1992; Zapf et al., 1989). Laut Zapf et al. (1989) lässt sich aufgrund der nicht vorhandenen bewussten Kontrolle nicht unterscheiden, ob sich Fehler bei der Planung, beim Gedächtnis/Monitoring oder beim Feedback ereignet haben. Es können deswegen auf dieser Ebene nur Fehler, die bei motorischen Bewegungen mit dem Eingabegerät (z.B. nicht vom Nutzer erkannte Tippfehler, Klickfehler) auftreten, eingeordnet werden (Zapf et al., 1989). Es handelt sich bei diesen Fehlern streng genommen um keine kognitiven Fehler bei der Assimilation, da solche zu einer Verarbeitung auf höheren Ebenen aufgrund des daraus entstehenden Bedarfs für kognitives Abkoppeln führen würden (z.B. bemerkt ein Nutzer einen Tippfehler im Nachhinein, handelt es sich um einen Unterlassensfehler und nicht mehr um einen reinen Bewegungsfehler) und man sie deswegen dann auch dort einordnen würde (siehe Teilabschnitt 2.2), sondern stattdessen lediglich um Fehler bei der motorischen Handlungsausführung (d.h. Ausrutscher: z.B. Nutzer drückt falsche Taste, bemerkt es aber nicht und arbeitet weiter). Wie bereits in Teilabschnitt 2.1.1 erklärt, lassen sich die motorische Effizienz bei der eigentlichen Handlungsausführung und die mentale Effizienz bei der kognitiven Informationsverarbeitung eindeutig voneinander abgrenzen, wobei nur letztere zur intuitiven Benutzung beiträgt (siehe Teilabschnitt 2.1.1) und die andere dem übergeordneten Konzept Usability zuordenbar ist (Hurtienne, 2011).

Nutzungsprobleme, die anhand der Fehlertaxonomie der sensomotorischen Ebene zugeordnet werden können, beeinträchtigen dementsprechend nicht die intuitive Benutzung des Systems (außer sie werden vom Nutzer erkannt und damit bewusst, wobei dann die Ursache höheren Ebenen zugeordnet werden kann). Im Zuge der formativen Evaluation intuitiver Benutzung müssen diese Probleme also auch nicht im Rahmen des menschzentrierten Gestaltungsprozesses beseitigt werden. Für die Sicherstellung der Gebrauchstauglichkeit eines Systems können diese Probleme jedoch relevant sein.

Bewegungsfehler Bewegungsfehler hängen mit den motorischen Fähigkeiten des Nutzers zusammen, die für die Ausführung einer Handlung benötigt werden. Bei der motorischen Umsetzung können Fehler auftreten, die in diese Kategorie eingeordnet werden können (Zandbergen, 2015; Zapf et al., 1992; Zapf et al., 1989).

3.2.4 Bilanzierend-summative Evaluation

Im Gegensatz zur formativen Evaluation erfolgt die bilanzierende bzw. summative Evaluation erst nach der Fertigstellung des Systems am Ende des menschzentrierten Gestaltungsprozesses DIN EN ISO 9241-210 (ISO, 2011) zur abschließenden oder vergleichenden Bewertung der Gesamtqualität (Sarodnick & Brau, 2006) und widmet sich daher den ersten beiden Evaluationszielen (d.h. vergleichende oder bewertende Evaluation). Diese Hauptart der Evaluation erfüllt demnach vor allem eine Kontroll- und Legitimationsfunktion, da es vorwiegend darum geht, das gestaltete System zusammenfassend, hinsichtlich bestimmter zuvor definierter Evaluationskriterien, vor dem Hintergrund eines bestimmten Verwendungskontextes (z.B. „Ist das fertige System besser als ein anderes System oder ist die Entwicklung des Systems wirklich abgeschlossen?“) oder übergreifend zur Sicherstellung der Güte zu beurteilen, um auf diese Weise Rechenschaft abzulegen (Döring & Bortz, 2016; Hegner, 2003; Sarodnick & Brau, 2006). Bei der summativen Evaluation spielen im Vergleich zur formativen Evaluation überwiegend quantitative Daten (z.B. Leistungsdaten wie die Anzahl erfolgreich gelöster Aufgaben) eine Rolle (Döring & Bortz, 2016; Hartson et al., 2001; Hegner, 2003). Summative Evaluationen benötigen für solche Bewertungen generell ein experimentelles Design, welches auf statistische Tests zurückgreift (Hartson et al., 2001). Da mithilfe einer bilanzierend-summativen Evaluation auch Erkenntnisse bezüglich der Güte des Evaluationsgegenstands (d.h. Meta-Evaluation) gewonnen werden können, die mit keiner unmittelbaren Verwendungsabsicht im Rahmen des menschzentrierten Gestaltungsprozesses in Verbindung stehen, sondern die Qualitätsstandards für jede dort eingesetzt Evaluation bilden (DeGEval, 2016), wird sich dieser Meta-Evaluation im folgenden Teilabschnitt gewidmet.

3.2.5 Meta-Evaluation in der HCI

Um eine Methode für die Evaluation von User Interfaces (speziell für 3D-CUIs im Rahmen des Projekts 3D-GUIde) bezüglich intuitiver Benutzung überhaupt einsetzen zu können (d.h. Evaluation bezüglich des Nutzens), muss zuvor deren Güte übergreifend bewertet werden (d.h. Evaluation bezüglich der Güte der Methode). Die dafür benötigten Evaluationskriterien lassen sich aus den folgenden vier Leitsätzen ableiten (DeGEval, 2016; Sanders, 2013):

Nützlichkeit Evaluationen sollen sich an zuvor definierten Evaluationszielen, Evaluationskriterien, sowie am generellen praktischen Informationsbedarf der Anwender orientieren (DeGEval, 2016).

Durchführbarkeit Evaluationen sollen realistisch, gut durchdacht, umsichtig und kosteneffizient durchgeführt werden können (DeGEval, 2016).

Fairness oder Korrektheit Evaluationen sollen rechtlich und ethisch korrekt ablaufen und dabei dem Wohlergehen an der Evaluation beteiligter Personen (d.h. evaluierte und evaluierende Personen) Aufmerksamkeit geschenkt werden (DeGEval, 2016).

Genauigkeit Evaluationen sollen wissenschaftlich korrekte und nachvollziehbare Ergebnisse zum jeweiligen Evaluationsgegenstand, den Evaluationskriterien und den Evaluationszielen hervorbringen, sowie diese auch angemessen vermitteln (DeGEval, 2016).

Die Genauigkeit einer Evaluation wird mithilfe sog. *formaler Hauptgütekriterien* sichergestellt, die als unverzichtbar angesehen werden, da eine Evaluation nur so wissenschaftlich tragfähige Messwerte liefern kann. Sie sichern damit die wissenschaftliche Güte des Evaluationsgegenstands (z.B. „Misst die Methode wirklich intuitive Benutzung?“) bzw. die *Qualität der Wissenschaft* (Döring & Bortz, 2016). Die Nützlichkeit, Durchführbarkeit und Fairness einer Evaluation wird hingegen üblicherweise mithilfe sog. *formaler Nebengütekriterien* sichergestellt, die lediglich „bedingte Anforderungen“ darstellen, deren Bedeutung abhängig von Evaluationszielen und Anwenderinteressen unterschiedlich sein kann. Sie sichern damit verschiedene Aspekte der praktischen Güte des Evaluationsgegenstands (z.B. „Ist die Methode wirklich zeitlich effizienter als gängige Methoden anwendbar?“) bzw. die *Qualität der Praxis* (Döring & Bortz, 2016).

Die praktische Güte einer Evaluationsmethode lässt sich jedoch erst nach der Sicherstellung der Grundvoraussetzungen der wissenschaftlichen Güte bewerten (DeGEval, 2016; Döring & Bortz, 2016; Lienert & Raatz, 1998), da mit dieser ja nur anwendungsspezifische Anforderungen, wie beispielsweise die im Projekt 3D-GUIde geforderte hohe zeitliche Anwendungseffizienz (siehe Kapitel 1), erfasst werden können. Die Überprüfung der Güte (d.h. wissenschaftliche und praktische Güte) einer Evaluation wird auch als *Meta-Evaluation* bezeichnet, da nun eine Evaluation selbst (d.h. Evaluationsmethode und damit verbundene Ergebnisse) den Evaluationsgegenstand einer Evaluation bildet (DeGEval, 2016). Im Rahmen dieser Arbeit soll der Begriff sowohl für die Sicherstellung der Güte neu entwickelter Methoden als auch für die Bewertung der Güte bereits bestehender Methoden genutzt werden. Da formative und summative Evaluationsmethoden einen unterschiedlichen Schwerpunkt haben und sich deswegen auch an einer unterschiedlichen Stelle im menschzentrierten Gestaltungsprozess wiederfinden, kommen hierfür auch unterschiedliche Prozesse der Meta-Evaluation zum Einsatz. Hierzu sei im Vorfeld darauf hingewiesen, dass zwar oft ähnliche Bezeichnungen für formative und summative Gütekriterien verwendet werden, sich aber hinter diesen Begriffen inhaltlich andere Kriterien verbergen. Dies ist der unterschiedlichen Ausrichtung von formativen und summativen Methoden geschuldet, wodurch sich auch die Auffassungen in der Literatur bezüglich der damit verbundenen Gütekriterien unterscheidet.

Zur *formalen Überprüfung der Güte formativer Evaluationsmethoden* (d.h. formative Meta-Evaluation) existiert in der Literatur bereits ein etablierter Kanon von Haupt- und Nebengütekriterien (z.B. Blandford, Green, Furniss, & Makri, 2008; Hartson et al., 2001; Koutsabasis, Spyrou, & Darzentas, 2007; Makri, Blandford, Cox, Attfield, & Warwick, 2011). Den Evaluationsgegenstand für eine formative Meta-Evaluation bildet die formative Methode selbst, sowie deren Evaluationsergebnisse, die die mithilfe einer Ursachenanalyse identifizierten realen Nutzungsprobleme darstellen (siehe Teilabschnitt 3.2.3). Auf deren Basis können anschließend im Zuge des bereits vorgestellten formellen vierstufigen Evaluationsprozesses (siehe Teilabschnitt 3.2.1) die Methode und die damit identifizierten

realen Nutzungsprobleme hinsichtlich einer Auswahl von formalen Haupt- und Nebengütekriterien (d.h. Evaluationskriterien) mit geeigneten Außenkriterien (z.B. Vergleich der Anzahl gefundener realer Probleme zwischen zwei verwandten Methoden) verglichen und basierend darauf über die allgemeine Güte der Methode entschieden werden. Die Auswahl der formalen Gütekriterien und Außenkriterien erfolgt dabei immer in Abhängigkeit vom Anwenderinteresse bzw. vom Nutzungskontext. Da praktische Güte wissenschaftliche Güte voraussetzt, sollten für die Sicherstellung der wissenschaftlichen Güte alle formalen formativen Hauptgütekriterien (d.h. Gründlichkeit, Gültigkeit, Zuverlässigkeit) geprüft und sich lediglich bei der Sicherstellung der praktischen Güte auf eine Auswahl von Nebengütekriterien beschränkt werden (z.B. Effizienz), die in Abhängigkeit vom Anwenderinteresse variiert (Blandford et al., 2008; Hartson et al., 2001; Koutsabasis et al., 2007; Makri et al., 2011). Da beispielsweise die Anforderung des 3D-GUIde Projekts an formative Methoden neben wissenschaftlicher Güte zusätzlich darin besteht, auf eine zeitlich anwendungseffiziente Evaluationsmethode für die Entwicklung von CUI-Interaktionspatterns zurückgreifen zu können, müssen die ausgewählten Nebengütekriterien auch diese Anforderung unterstützen.

Zur *formalen Überprüfung der Güte summativer Evaluationsmethoden* existiert in der Literatur ebenfalls ein etablierter Kanon von Haupt- und Nebengütekriterien (z.B. Döring & Bortz, 2016; Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007). Den Evaluationsgegenstand für eine summative Meta-Evaluation bildet die summative Methode selbst, sowie deren Evaluationsergebnisse, die die Ausprägung eines gemessenen latenten (d.h. nicht direkt beobachtbaren) Merkmals (z.B. intuitive Benutzung) widerspiegeln (siehe Teilabschnitt 3.2.4). Auf deren Basis können anschließend im Zuge des bereits vorgestellten vierstufigen Evaluationsprozesses (siehe Teilabschnitt 3.2.1) die Methode und die Ausprägung des gemessenen latenten Merkmals hinsichtlich einer Auswahl von formalen Haupt- und Nebengütekriterien (d.h. Evaluationskriterien) mit geeigneten Außenkriterien (z.B. Vergleich des gemessenen Ausmaßes intuitiver Benutzung zwischen zwei verwandten Methoden) verglichen und auf dieser Basis über die Güte der Methode entschieden werden. Die Auswahl der formalen Gütekriterien und Außenkriterien erfolgt dabei immer in Abhängigkeit vom Anwenderinteresse bzw. vom vorliegenden Nutzungskontext. Da praktische Güte wissenschaftliche Güte voraussetzt, sollten für die Sicherstellung der wissenschaftlichen Güte alle formalen Hauptgütekriterien (d.h. Objektivität, Validität, Reliabilität) geprüft und sich lediglich bei der Sicherstellung der praktischen Güte auf eine Auswahl von Nebengütekriterien beschränkt werden (z.B. Effizienz), die in Abhängigkeit vom Anwenderinteresse variiert (Döring & Bortz, 2016; Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007). Da die Anforderung des 3D-GUIde Projekts an summative Methoden neben wissenschaftlicher Güte darin besteht, auf eine zeitlich anwendungseffiziente Evaluationsmethode für die Entwicklung von 3D-CUI-Interaktionspatterns zurückgreifen zu können, müssen die ausgewählten Nebengütekriterien auch diese Anforderung unterstützen.

Problematisch an der beschriebenen formalen Vorgehensweise bei der formativen und summativen Meta-Evaluation (siehe Teilabschnitt 3.2.1), bei der eine Auswahl von formalen Gütekriterien mit vorab festgelegten Außenkriterien verglichen wird, ist es, dass bereits geeignete Außenkriterien für diesen Vergleich existieren müssen, um auf diese Weise einer neuen untersuchten Methode (d.h. Evaluationsgegenstand der Meta-Evaluation) Güte (d.h. wissenschaftliche und praktische Güte) attestieren oder eine bereits existierende Methode unter diesem Gesichtspunkt für ein Forschungsvorhaben auswählen zu können.

Sowohl im formativen (siehe Hartson et al., 2001) als auch im summativen (siehe Döring & Bortz, 2016; Lienert & Raatz, 1998) Bereich lässt sich ein echtes Außenkriterium, das semantische und theoretische Ähnlichkeit zur untersuchten Evaluationsmethode aufweist und zusätzlich von höherem Status (z.B. Sterberate als echtes Außenkriterium für Alkoholkonsum) als diese selbst ist, oftmals nicht adäquat auf objektive Art und Weise bestimmen. Es wird in solchen Fällen daher ein echtes Außenkriterium anhand mehrerer semantisch und theoretisch ähnlicher Evaluationsmethoden (d.h. Quasi-Außenkriterium), die jedoch nicht von höherem Status als die untersuchte Evaluationsmethode selbst sind (z.B. ein Fragebogen zum Alkoholkonsum wird mithilfe mehrerer Fragebögen zum Alkoholkonsum validiert), approximiert. Echte Außenkriterien und Quasi-Außenkriterien sollen im Rahmen dieser Arbeit zusammenfassend als Außenkriterien bezeichnet werden.

Bei diesem „Workaround“ kann es leicht zu einem Problem kommen, wenn für eine Meta-Evaluation auf Quasi-Außenkriterien zurückgegriffen wird, die selbst nicht ausgiebig bezüglich der in Abhängigkeit des Anwenderinteresses gewählten formalen Gütekriterien (z.B. wissenschaftliche Güte nur teilweise bestätigt) evaluiert sind. Im Rahmen dieser Arbeit sollen deswegen, sofern es möglich ist, nur Quasi-Außenkriterien für die Approximation eines echten Außenkriteriums genutzt werden, deren wissenschaftliche Güte bereits formal anhand aller vorhandenen Hauptgütekriterien sichergestellt wurde. Quasi-Außenkriterien, die von allen verfügbaren Quasi-Außenkriterien im Forschungsfeld zu intuitiver Benutzung die höchste formal nachgewiesene wissenschaftliche Güte als eine formative oder summative Evaluationsmethode aufweisen, werden im Rahmen dieser Arbeit auch als *Benchmarks* bezeichnet (d.h. formativer Benchmark und summativer Benchmark). Es wurde sich für diese Bezeichnung entschieden, da diese Evaluationsmethoden den aktuellen Maßstab im Forschungsfeld zu intuitiver Benutzung bilden und somit am besten für die Meta-Evaluation der in Abhängigkeit des Anwenderinteresses gewählten Nebengütekriterien geeignet sind. Bezogen auf das Projekt 3D-GUIde ist bei solchen Benchmarks die wissenschaftliche Güte mit hoher Wahrscheinlichkeit formal sichergestellt, weswegen sich auch nur hier die im Projekt speziell geforderte zeitliche Anwendungseffizienz anhand ausgewählter Nebengütekriterien verlässlich beurteilen lässt. Auf diese Weise läuft man nicht Gefahr, eine zeitlich anwendungseffiziente Methode zu verwenden, deren wissenschaftliche Güte aber noch nicht formal sichergestellt wurde. Demzufolge können nur solche Benchmarks aktuell für die formative (d.h. formativer Benchmark) und summative (d.h. summativer Benchmark) Evaluation von 3D-CUI-Interaktionslösungen im Rahmen des Projekts 3D-GUIde eingesetzt werden.

Darüber hinaus kann es auch vorkommen, dass nur ein Quasi-Außenkriterium im Feld existiert, um damit ein echtes Außenkriterium approximieren zu können. Eine derartige Problematik lässt sich im Forschungsbereich zu intuitiver Benutzung bei der Berücksichtigung eines aktuellen Reviews von Blackler et al. (2018) feststellen, da dort eine Vielzahl von Methoden oft nur eingeschränkt bezüglich ihrer wissenschaftlichen Güte evaluiert sind (z.B. wissenschaftliche Güte wurde nicht formal durch alle Hauptgütekriterien sichergestellt) und sich im Forschungsfeld verfügbare Methoden dadurch auch automatisch für den Einsatz im Projekt 3D-GUIde als Benchmark disqualifizieren. Betrachtet man das angesprochene Review genauer, propagiert dieses Review zwar, dass Evaluatoren über einen reichen Methodenfundus verfügen, thematisiert jedoch an keiner Stelle die formale Sicherstellung der wissenschaftlichen Güte der vorgestellten Methoden. Auch vorherige Reviews (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015) des Forschungsfeldes

zu intuitiver Benutzung geben hierzu keinerlei Auskunft. Schaut man sich aufgrund dieser mangelnden Informationen, die im aktuellsten Review referenzierten Veröffentlichungen, sowie aktuelle Veröffentlichungen der Methoden direkt an, wird die wissenschaftliche Güte kaum oder gar nicht diskutiert (z.B. Blackler, 2006; Brandenburg & Sachse, 2012; Desai, Blackler, & Popovic, 2015; Lawry, 2012; Macaranas et al., 2015; McEwan, Blackler, Johnson, & Wyeth, 2014; Reddy, 2012) und nur bestimmte Autoren thematisieren diese explizit bei ihren Methoden (z.B. Horn, 2008; McAran, 2018; Naumann & Hurtienne, 2010; Reinhardt et al., 2018; Ullrich, 2014; Ullrich & Diefenbach, 2010b).

Betrachtet man unter diesem Gesichtspunkt auch aktuelle Reviews aus dem Bereich Usability, dem intuitiver Benutzung übergeordneten Konzept (Hurtienne, 2011), zeichnet sich ein ähnliches Bild ab. Laut einem aktuellen Review von Stanton (2016) ist der Methodenfundus hier noch riesiger, aber es herrschen ebenfalls große Inkonsistenzen bei der formalen Sicherstellung der wissenschaftlichen Güte der im Review referenzierten Methoden. Statt die wissenschaftliche Güte anhand eines Vergleichs der Methode mit Außenkriterien anhand von formalen Hauptgütekriterien abzusichern, ein Vorgehen was sich in anderen verwandten Forschungsgebieten wie der Psychologie schon lange etabliert hat, wird diese formale Evaluation oftmals vernachlässigt, da man im HCI-Bereich jahrelang implizit davon ausgegangen ist, dass, wenn Experten Evaluationsmethoden ordnungsgemäß anwenden, diese auch ohne eine explizite Sicherstellung der wissenschaftlichen Güte inhaltlich valide Ergebnisse produzieren (Kanis, 2014; Stanton, 2016; Stanton & Young, 1998). Wenn die Güte dennoch einmal Thema ist, werden dabei in diesem Zusammenhang vorwiegend praktische Aspekte fokussiert. Viele Forscher stützen hierbei ihre Argumentation auf eher nicht formale Kriterien, die ohne einen expliziten Vergleich mit Außenkriterien auskommen und mehr dem Anspruch entwachsen sind, Forschern bei der Entscheidung für eine bestimmte Methode für einen bestimmten Anwendungsfall zu dienen (Kanis, 2014; Stanton, 2016; Stanton & Young, 1998). Als Beispiele für auf diese Weise validierte Evaluationsmethoden können an dieser Stelle die Arbeiten von Jeffries und Desurvire (1992), Dutt, Johnson und Johnson (1994), Molich und Dumas (2008) und Simeral und Branaghan (1997) genannt werden. Nur wenige Arbeiten stellen die wissenschaftliche Güte ihrer Methoden im Usability-Bereich zusätzlich mithilfe von formalen Hauptgütekriterien sicher (z.B. Agarwal & Venkatesh, 2002; Makri et al., 2011; Panach, Condori-Fernández, Valverde, Aquino, & Pastor, 2008; Yen, Sousa, & Bakken, 2014).

In diesem Zusammenhang ist auch der kontroverse Artikel „Damaged Merchandise“ von Gray und Salzman (1998) zu nennen, der allgemein methodische Kritik bezüglich der Meta-Evaluation von Evaluationsmethoden im gesamten Forschungsfeld der HCI übt und dabei die generelle Güte der untersuchten Verfahren anzweifelt. Dieser Artikel und die damit ausgelöste Debatte führte bei vielen Forschern zu der Ansicht, dass es eigentlich unmöglich ist, eine methodisch saubere Evaluationsmethode zu entwickeln (Blandford et al., 2008). Zusätzlich führte die Kontroverse um den Artikel von Gray und Salzman (1998) dazu, dass immer weniger Meta-Evaluationen für Evaluationsmethoden veröffentlicht wurden, was nach Blandford et al. (2008) zu einer hohen Übergeneralisierung führt und ein tiefgreifendes Verständnis der Stärken und Schwächen von Evaluationsmethoden verhindert. Nicht nur Evaluationsmethoden kämpfen mit dieser Problematik, sondern eigentlich alle Methoden in der HCI (Kanis, 2014; Stanton, 2016; Stanton & Young, 1998).

Der Grund dafür besteht darin, dass entwickelte Methoden (z.B. auch Gestaltungsmethoden) in der HCI viel zu oft für Problemstellungen und Umgebungen verwendet werden (d.h. Nutzungskontext), für die sie ursprünglich gar nicht vorgesehen waren und evaluiert wurden (Kanis, 2014; Stanton, 2016; Stanton & Young, 1998). Für eine vertiefte Auseinandersetzung auf die damit verbundene Debatte der Übergeneralisierung von Methoden im Rahmen von menschzentrierten Gestaltungsprozessen wird an dieser Stelle auf Kaptein und Robertson (2012), Bias, Kortum, Sauro und Gillan (2013) und Tractinsky (2018) verwiesen. Da sich intuitive Benutzung in der mentalen Beanspruchung des Nutzers widerspiegelt (siehe Kapitel 2), werden viele Methoden innerhalb des Forschungsfelds oftmals anhand der folgenden Gütekriterien „evaluiert“, die im HCI-Forschungsbereich zu mentaler Beanspruchung rege Anwendung finden (z.B. Cain, 2007; Eggemeier, 1988; Longo, 2017; O'Donnell & Eggemeier, 1986):

Sensitivität Evaluationsmethoden sollen tatsächliche Unterschiede eines gemessenen Merkmals (z.B. intuitive Benutzung) wahrnehmen können. Laut Beatty (1982) gilt eine Methode als sensitiv, wenn sie (1) interindividuelle Fähigkeitsunterschiede, (2) Schwierigkeitsunterschiede zwischen strukturell identischen Aufgaben, und (3) Schwierigkeitsunterschiede zwischen strukturell unterschiedlichen Aufgaben abbilden kann (Cain, 2007).

Spezifität Evaluationsmethoden sollen nur Unterschiede abbilden können, die auf Änderungen eines bestimmten Merkmals (z.B. intuitive Benutzung) und nicht auf Änderungen eines anderen bestimmten Merkmals (z.B. physischer Effizienz) zurückzuführen sind (Cain, 2007).

Diagnostizität Evaluationsmethoden sollen nicht nur Unterschiede eines gemessenen Merkmals (z.B. intuitive Benutzung) feststellen können, sondern auch die Ursachen die dafür verantwortlich sind (Cain, 2007).

Reliabilität Evaluationsmethoden sollen das Merkmal (z.B. intuitive Benutzung) konsistent erfassen können (Cain, 2007).

Intrusion Evaluationsmethoden sollen nicht selbst die Ausprägung des untersuchten Merkmals (z.B. intuitive Benutzung) stören oder behindern (Cain, 2007).

Umsetzungsanforderungen Evaluationsmethoden sollen für ihre Anwendung nur geringe Anforderungen an Ressourcen wie beispielsweise Durchführungszeit, Training, Anzahl benötigter Evaluatoren, Hard- und Software stellen. Des Weiteren wird mit dieser Dimension auch das Verhältnis von Kosten, Schwierigkeit der Datenauswertung und entsprechendem Nutzen beschrieben (Cain, 2007).

Benutzerakzeptanz Evaluationsmethoden sollen für den Anwender zumindest augenscheinvalide sein (d.h. für Laien plausibel sein) und eine gewisse Nützlichkeit im jeweiligen Nutzungskontext aufweisen (Cain, 2007).

Mithilfe der ersten fünf Kriterien sichern Forscher üblicherweise auf nicht formale Art und Weise ohne Vergleich mit einem Außenkriterium die wissenschaftliche Güte ihrer Methode ab. Die restlichen beiden Kriterien dienen der Sicherstellung verschiedener Aspekte der praktischen Güte (Hancock & Matthews, 2019; Matthews, Reinerman-Jones, Barber, & Abich IV, 2015). Matthews et al. (2015) und Hancock und Matthews (2019) kritisieren bei diesem nicht formalen Vorgehen, dass mithilfe dieser nicht formalen Gütekriterien, die wissenschaftliche Güte einer Evaluationsmethode unzureichend sichergestellt wird, da die ver-

wendeten Gütekriterien keinen Vergleich mit Außenkriterien vorsehen, so wie es eigentlich streng genommen in einem formalen Evaluationsprozess gefordert ist (siehe Teilabschnitt 3.2.1). Außerdem werden laut Matthews et al. (2015) und Hancock und Matthews (2019) die Unterschiede von summativen und formativen Evaluationen in den nicht formalen Gütekriterien unausgewogen berücksichtigt (z.B. formative Aspekte werden eigentlich nur durch die Diagnostizität thematisiert). Im weiteren Verlauf dieser Arbeit wird des Öfteren auf diese nicht formalen Gütekriterien Bezug genommen und sie werden zu diesem Zweck immer als *nicht formale Gütekriterien* bezeichnet, um sie von den *formalen Gütekriterien* abzugrenzen. Evaluationsmethoden, die lediglich anhand dieser nicht formalen Kriterien ohne Vergleich mit einem Außenkriterium (d.h. Quasi-Kriterium oder echtes Außenkriterium) „evaluiert“ worden sind, können somit keine geeigneten Quasi-Außenkriterien zur Approximation eines echten Außenkriteriums und damit auch keinen Benchmark innerhalb des Forschungsfeldes zu intuitiver Benutzung darstellen (siehe Hancock & Matthews, 2019; Matthews et al., 2015).

In Konsequenz führt diese unzureichende Sicherstellung der wissenschaftlichen Güte dazu, dass viele Evaluationsmethoden in den für intuitive Benutzung relevanten Forschungsgebieten (z.B. Forschung zu Usability, mentaler Beanspruchung und intuitiver Benutzung selbst) streng genommen nicht zur Approximation eines echten Außenkriteriums für eine Meta-Evaluation einer neu entwickelten Methode genutzt und damit auch nicht als Benchmark für die Evaluation intuitiver Benutzung im Projekt 3D-GUIde eingesetzt werden können. Nutzt man für diesen Zweck dennoch Evaluationsmethoden, deren wissenschaftliche Güte nicht formal sichergestellt wurde, führt dies zu einem Zirkelschluss, weil man nicht genau weiß, ob die evaluierte Methode wirklich wissenschaftlich valide Ergebnisse produzieren kann, oder diese Einschätzung nur auf die geringe wissenschaftliche Güte der gewählten Vergleichsmethoden zurückzuführen ist (Lienert & Raatz, 1998). Dementsprechend wird in einschlägiger Literatur immer dazu geraten für eine Meta-Evaluation Methoden von hoher wissenschaftlicher Güte als Quasi-Außenkriterien für die Approximation eines echten Außenkriteriums einzusetzen (Baumgartner, 1999; Döring & Bortz, 2016; Lienert & Raatz, 1998).

Zusammenfassend lässt sich unter Berücksichtigung der Rahmenbedingungen des Anwendungsprojekts 3D-GUIde festhalten, dass trotz der schlechten Qualität der Meta-Evaluation im Forschungsfeld zu intuitiver Benutzung, nur Methoden als Quasi-Außenkriterien für eine die wissenschaftliche Güte formal sicherstellende Meta-Evaluation genutzt werden sollten, deren wissenschaftliche Güte selbst formal nachgewiesen wurde. Stellen diese Methoden zusätzlich einen Benchmark im Forschungsfeld dar, können sie im Rahmen des Projekts bei der Evaluation von 3D-CUI-Interaktionslösungen bereits formativ und summativ eingesetzt, sowie daran auch die wissenschaftliche Güte und die zeitliche Anwendungseffizienz einer neuen Methode beurteilt werden. Sollten neben diesen Benchmarks keine weiteren Quasi-Außenkriterien im Feld verfügbar sein, deren wissenschaftliche Güte bereits formal nachgewiesen ist, kann der jeweilige Benchmark allein für die Approximation eines echten Außenkriteriums im Rahmen einer Meta-Evaluation genutzt werden. Nachfolgend werden nun formale formative und summativen Gütekriterien vorstellt mit deren Hilfe neue Evaluationsmethoden im Vergleich mit Außenkriterien formal evaluiert werden können.

3.3 Formale Gütekriterien der formativen Evaluation

Um die Güte einer formativen Evaluation sicherzustellen, kann als Standardkanon auf etablierte formale Gütekriterien zurückgegriffen werden, die von einer Reihe von Forschern im Usability-Bereich kontinuierlich angewendet werden (z.B. Blandford et al., 2008; Hartson et al., 2001; Hasan, Morris, & Probets, 2012; Koutsabasis et al., 2007; Lanzilotti, Ardito, Costabile, & De Angeli, 2011). Als formale Hauptgütekriterien kommen dabei *Gründlichkeit*, *Gültigkeit* und *Zuverlässigkeit* zum Einsatz, die als Mindestvoraussetzungen für eine neu entwickelte formative Evaluationsmethode fungieren und ihre wissenschaftliche Güte sichern können (Hartson et al., 2001). Dabei setzt Gültigkeit immer Gründlichkeit voraus, welche beide die zentralen Hauptgütekriterien der formativen Evaluation im Rahmen des menschzentrierten Gestaltungsprozesses darstellen. Zuverlässigkeit ist im Vergleich dazu von geringerer Priorität (Hartson et al., 2001; Makri et al., 2011). Unter Berücksichtigung der Anforderungen aus dem Projekt 3D-GUIde sollten alle formalen Hauptgütekriterien erfüllt sein, um einer formativen Evaluationsmethode wissenschaftliche Güte attestieren zu können.

Neben diesen zur Sicherstellung der wissenschaftlichen Güte notwendigen formalen Hauptgütekriterien existieren auch noch eine Reihe von formalen Nebengütekriterien, die sich in ihrer Bedeutung abhängig von Anwenderinteressen unterscheiden können und der Sicherstellung der praktischen Güte der formativen Evaluationsmethode bzw. als praktische Entscheidungshilfe dienen (siehe Blandford et al., 2008; Hartson et al., 2001; Hasan et al., 2012; Koutsabasis et al., 2007; Lanzilotti et al., 2011). Unter Berücksichtigung der Anforderungen aus dem Projekt 3D-GUIde muss lediglich das Nebengütekriterium Effizienz (d.h. zeitliche Anwendungseffizienz) erfüllt sein, um einer formativen Evaluationsmethode praktische Güte aus Anwendungsprojektsicht attestieren zu können. Alle formalen formativen Gütekriterien werden nun nachfolgend im Detail vorgestellt.

3.3.1 Hauptgütekriterium Gründlichkeit

Das formale formative Hauptgütekriterium *Gründlichkeit* beschreibt den Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der vorhandenen realen Nutzungsprobleme im untersuchten System (Hartson et al., 2001). Mit diesem Hauptgütekriterium soll die prinzipielle Anforderung an eine formative Evaluationsmethode erfüllt werden, möglichst viele real existierende Nutzungsprobleme bei der Systemnutzung aufspüren zu können (Hartson et al., 2001). Findet eine zu validierende Evaluationsmethode beispielsweise nur 10 von 20 realen Problemen, die sich eigentlich in dem untersuchten System befinden, besitzt sie eine Gründlichkeit von 50 Prozent. Um diese Hauptgütekriterium anwenden zu können, muss laut Blandford et al. (2008) und Hartson et al. (2001) zunächst geklärt werden, was genau unter einem realen Nutzungsproblem zu verstehen ist, wie es bereits in Teilabschnitt 3.2.3 im Hinblick auf intuitive Benutzung diskutiert wurde. Jedoch lässt sich ein reales Nutzungsproblem auch anders definieren. Beispielsweise lassen sich auch alle als katastrophal bewerteten Nutzungsprobleme als echt auffassen, nur die Probleme die eine bestimmte Anzahl an Nutzern hatten oder nur die Probleme, die die Effektivität der Nutzung verhindert haben usw.

Laut Hartson et al. (2001) handelt es sich um ein reales Nutzungsproblem, wenn dieses in einem echten Nutzungskontext auftreten kann und einen Einfluss auf die Gebrauchstauglichkeit hat. Da im Rahmen dieser Arbeit mit intuitiver Benutzung ein bestimmter Teilaspekt von Gebrauchstauglichkeit betrachtet wird, werden nur Nutzungsprobleme als real betrachtet, die zu erhöhter mentaler Beanspruchung des Nutzers führen und damit eine intuitive Benutzung des Nutzers beeinträchtigen können. Laut der in Teilabschnitt 3.2.3 vorgestellten Ursachenanalyse handelt es sich nur bei Nutzungsproblemen die der intellektuellen Regulationsebene (d.h. Denkfehler, Merk- und Vergessensfehler, Urteilsfehler) und der perzeptiv-begrifflichen Regulationsebene (d.h. Gewohnheitsfehler, Unterlassensfehler, Erkennensfehler) zugeordnet werden können, um echte Probleme im Sinne intuitiver Benutzung, da diese das kognitive Abkoppeln und damit die mentale Beanspruchung beeinflussen. Fehler auf der sensomotorischen Regulationsebene (d.h. Bewegungsfehler) beeinflussen lediglich die motorische Handlungsausführung und Ineffizienzen (d.h. Lösungen, die zwar zum Ziel führen, aber umständlich sind) beeinflussen lediglich die zeitliche Effizienz, wirken sich daher nicht negativ auf die mentale Beanspruchung des Nutzers aus und sind somit nicht als echte Nutzungsprobleme zu klassifizieren.

Des Weiteren merken Blandford et al. (2008) und Hartson et al. (2001) zu diesem formalen Hauptgütekriterium im Allgemeinen an, dass es sehr schwierig und nahezu unmöglich ist, die Gesamtzahl an vorhandenen realen Problemen objektiv anhand eines echten Außenkriteriums vor der Meta-Evaluation zu bestimmen. Wie bereits in Teilabschnitt 3.2.5 erwähnt, wird deswegen das echte Außenkriterium mithilfe von Quasi-Außenkriterien oder zumindest des im Forschungsfeld existierenden formativen Benchmarks approximiert. Üblicherweise wird die Gesamtanzahl aller realen Nutzungsprobleme dadurch bestimmt, dass alle realen Probleme, die durch die verschiedenen Quasi-Außenkriterien gefunden werden, aggregiert werden.

3.3.2 Hauptgütekriterium Gültigkeit

Das formale formative Hauptgütekriterium *Gültigkeit* beschreibt den Anteil, der durch die formative Evaluationsmethode gefundenen realen Nutzungsprobleme an der Gesamtzahl der von dieser formativen Evaluationsmethode gefundenen Probleme (Hartson et al., 2001). Auf diese Weise soll sie dem Anspruch einer formativen Evaluationsmethode gerecht werden, damit möglichst nur reale Probleme zu finden (Hartson et al., 2001). Wie bereits erwähnt, kann mithilfe der in Teilabschnitt 3.2.3 vorgestellten Ursachenanalyse über die Echtheit gefundener Probleme in Bezug intuitiver Benutzung entschieden werden. Eine formative Evaluationsmethode, die 20 Probleme im untersuchten System findet, von denen nur fünf mithilfe der Ursachenanalyse als echt eingestuft werden können, hat in diesem Fall einen Gültigkeitsquotienten von 25 Prozent. Blandford et al. (2008) erklären, dass durch hohe Gültigkeit einer formativen Evaluationsmethode sichergestellt wird, dass Falschmeldungen (d.h. Aspekte, die irrtümlicherweise als Nutzungsprobleme bei der Systemnutzung identifiziert wurden) und Fehlschlüsse (d.h. mangelnde Gründlichkeit durch Übersehen von tatsächlichen Nutzungsproblemen bei der Systemnutzung durch die Evaluationsmethode, welche eigentlich von ihr erfasst werden müssten) minimiert werden und so das Nutzerverhalten besser vorhergesagt werden kann. Formative Evaluationsmethoden mit geringer Gültigkeit finden laut Hartson et al. (2001) üblicherweise eine große Anzahl

an Nutzungsproblemen, bei denen es sich nicht um tatsächliche Probleme handelt, und verschwenden dadurch viele Ressourcen bei einer Evaluation (z.B. Zeit und Kosten für Evaluation, Berichterstattung und Fehlerbeseitigung).

3.3.3 Hauptgütekriterium Zuverlässigkeit

Die *Zuverlässigkeit* beschreibt als letztes formales formatives Hauptgütekriterium, inwieweit die gesammelten realen Nutzungsprobleme, der zu untersuchenden Evaluationsmethode über mehrere Evaluatoren hinweg konsistent sind und damit, inwiefern die durch die Evaluationsmethode gefundenen Nutzungsprobleme unabhängig vom Evaluator sind. Zur Bestimmung der Zuverlässigkeit wird laut Hartson et al. (2001) üblicherweise die Beurteilerübereinstimmung der Evaluatoren anhand eines durchschnittlichen Korrelationskoeffizienten (Nielsen, 1994) oder durch Einschätzen der Konkordanz (z.B. bei zwei Evaluatoren: Cohens κ, siehe J. Cohen, 1960) ermittelt.

3.3.4 Nebengütekriterien

Neben den eben vorgestellten formalen Hauptgütekriterien, die über die wissenschaftliche Güte einer formativen Evaluationsmethode entscheiden, können im Nachgang noch Nebengütekriterien hinzugezogen werden, sofern die Hauptgütekriterien als erfüllt gelten (Hartson et al., 2001; Makri et al., 2011). Diese formalen Nebengütekriterien dienen der Sicherstellung der praktischen Güte der formativen Evaluationsmethode und deren Auswahl kann, wie bereits angesprochen, in Abhängigkeit vom konkreten Anwenderinteresse variieren. Im Gegensatz zu den vorgestellten Hauptgütekriterien sind die Nebengütekriterien eher von qualitativer Natur und daher weniger leicht quantifizierbar (Makri et al., 2011). Mithilfe der folgenden formalen Nebengütekriterien, denen bei unterschiedlichen Sammlungen von Nebengütekriterien eine besonders hohe Bedeutung zugeordnet wird (z.B. Blandford et al., 2008; Hartson et al., 2001; Hasan et al., 2012; Koutsabasis et al., 2007; Lanzilotti et al., 2011), kann die im Rahmen des Projekts 3D-GUIde geforderte zeitliche Anwendungseffizienz im Vergleich mit dem formativen Benchmark innerhalb des Forschungsfeldes zu intuitiver Benutzung sichergestellt werden:

Effektivität Effektivität ist das Produkt der Hauptgütekriterien *Gründlichkeit* und *Gültigkeit*. Dieses Nebengütekriterium wurde von Hartson et al. (2001) vorwiegend aus praktischen Gründen als Nebengütekriterium vorgeschlagen, da weder Gründlichkeit noch Gültigkeit alleine den praktischen Nutzen einer formativen Evaluationsmethode sichern können. Zum Beispiel schließt eine hohe Gründlichkeit nicht aus, dass sich auch nicht echte Probleme einschleichen und eine hohe Gültigkeit garantiert nicht, dass echte Probleme vergessen wurden (Hartson et al., 2001). Die Sicherstellung beider Kriterien ist für Hartson et al. (2001) entscheidend, wird aber von Blandford et al. (2008) aufgrund des hohen Aufwands (z.B. Bestimmung der Gesamtzahl der real vorhandenen Nutzungsprobleme in einem System) in der Praxis als besonders schwer eingestuft.

Effizienz Dieses Kriterium wird von Hartson et al. (2001) als die Kombination von Kosten (d.h. Kosten für die Einarbeitung und die Anwendung der untersuchten formativen Evaluationsmethode) und Effektivität der formativen Evaluationsmethode (d.h.

wissenschaftliche Güte durch Gewährleistung der Effektivität) definiert. Um in der Praxis über den Einsatz einer Methode entscheiden zu können, müssen verschiedene Methoden bezüglich Kosten und Effektivität abgewogen werden. Hartson et al. (2001) empfehlen dazu das Bilden eines Kosten-Nutzen-Quotienten.

Neben diesen beiden grundlegenden formalen Nebengütekriterien existieren in der Literatur noch weitere Nebengütekriterien, die nicht nur die prinzipielle Nützlichkeit der Methode betrachten, sondern sich speziell dem Erfolg einer tatsächlichen Einführung in die Praxis und der damit verbundenen erfolgreichen Verwertung der Evaluationsergebnisse im menschzentrierten Gestaltungsprozess widmen (Blandford et al., 2008; Hartson et al., 2001; Makri et al., 2011). Da mit diesen Kriterien nicht die zeitliche Anwendungseffizienz der formativen Methode im Rahmen des Projekts 3D-GUIde beurteilt werden kann, sollen diese Nebengütekriterien in der vorliegenden Arbeit keine explizite Anwendung finden, aber aufgrund der Vollständigkeit erwähnt und implizit mitbedacht werden:

Geltungsbereich Dieses Nebengütekriterium gibt Auskunft darüber, welcher Arten von Nutzungsproblemen bei der Systemnutzung mithilfe der formativen Evaluationsmethode entdeckt werden können (Makri et al., 2011).

Abgeleitete Erkenntnisse Dieses Nebengütekriterium beschreibt, inwiefern die Methode Erkenntnisse für etwaige Neugestaltungen des untersuchten Systems liefert. Dabei hängt das Kriterium sehr stark mit dem Geltungsbereich zusammen, da Methoden mit verschiedenen Geltungsbereichen auch mit höherer Wahrscheinlichkeit zu verschiedenen Typen an Erkenntnissen führen. Darüber hinaus steht das Kriterium auch mit der Downstream Utility in Verbindung (Makri et al., 2011).

Downstream Utility Dieses Nebengütekriterium beschreibt die Nützlichkeit der formativen Evaluationsmethode innerhalb des gesamten iterativen menschzentrierten Gestaltungsprozesses nachdem die Problembeschreibungen durch die neue Methode gesammelt wurden (Hartson et al., 2001). Dieser Aspekt ist besonders wichtig, da es für den kommerziellen Erfolg bzw. die allgemeine Verbesserung eines Systems entscheidend ist, inwieweit Evaluationsergebnisse in nützliche Verbesserungsvorschläge überführt werden können (Hartson et al., 2001; Makri et al., 2011).

Praxistauglichkeit Dieses Nebengütekriterium beschreibt die zu erfüllenden Voraussetzungen, um die formative Evaluationsmethode in der Praxis einsetzen zu können. Dies beinhaltet die Prüfung von Einschränkungen und Zielkonflikten, die zur Sicherstellung der Downstream Utility benötigt werden (Makri et al., 2011).

Evaluatoraktivitäten Dieses Nebengütekriterium beschreibt die Tätigkeiten eines Evaluators bei der Anwendung der untersuchten Evaluationsmethode in einer strukturierten Art und Weise. Bezüglich dieses Kriteriums herrscht aktuell in der Forschungsgemeinde keine Einigkeit, wie sich dieses Kriterium strukturiert erfassen lässt (Makri et al., 2011). Es existieren jedoch bereits einige wissenschaftliche Arbeiten, die sich an einem strukturierten Vorgehen versuchen (siehe Blandford et al., 2008).

Überzeugungskraft Dieses Nebengütekriterium beschreibt die Fähigkeit eines Evaluators mithilfe der durch die Methode erhaltenen Evaluationsergebnisse einen Entwickler von einer Systemänderung überzeugen zu können (Blandford et al., 2008; Makri et al., 2011).

Usability Dieses Nebengütekriterium beschreibt die Usability der Methode selbst und stellt somit die Einfachheit der Anwendung der Methode für die intendierte Zielgruppe an Evaluatoren dar (Makri et al., 2011). Blandford et al. (2008) sehen bei der Diskussion der allgemeinen Usability einer Methode insbesondere die Kriterien Erlernbarkeit, Kosteneffizienz und Integrierbarkeit in die Designpraxis im Vordergrund.

Erlernbarkeit Dieses Nebengütekriterium beschreibt die Erlernbarkeit der untersuchten formativen Evaluationsmethode, welche einen entscheidenden Faktor bei der Einführung der Methode in die Praxis darstellt. Je einfacher und schneller ein Evaluator mit der Methode arbeiten kann, umso wahrscheinlicher wird sich diese auch in der Praxis durchsetzen können (Blandford et al., 2008; Makri et al., 2011).

3.4 Formale Gütekriterien der summativen Evaluation

Um die Güte einer summativen Evaluation formal sicherzustellen, kann als Standardkanon auf psychometrische Gütekriterien zurückgegriffen werden, da die summative Evaluation ähnlich wie klassische psychologische Tests das Ziel verfolgt, die Ausprägung eines latenten psychologischen Merkmals (z.B. intuitive Benutzung) messbar zu machen, um bestimmte wissenschaftliche Aussagen oder praktische Entscheidungen treffen zu können (Döring & Bortz, 2016). Die wissenschaftliche Güte wird dabei durch die klassischen drei Hauptgütekriterien *Objektivität*, *Reliabilität* und *Validität* formal abgesichert. Diese stellen die Mindestanforderungen für eine neu entwickelte summative Evaluationsmethode dar (Döring & Bortz, 2016; Lienert & Raatz, 1998). Dabei bildet Objektivität die Voraussetzung für Reliabilität und Reliabilität wiederum die Voraussetzung für Validität (Bühner, 2011). Unter Berücksichtigung der Anforderungen aus dem Projekt 3D-GUIde sollten alle formalen Hauptgütekriterien erfüllt sein, um einer summativen Evaluationsmethode wissenschaftliche Güte attestieren zu können.

Neben diesen zur Sicherstellung der wissenschaftlichen Güte notwendigen formalen Hauptgütekriterien existieren noch eine Reihe von formalen Nebengütekriterien, die sich in ihrer Bedeutung abhängig von Anwenderinteressen unterscheiden können, und der Sicherstellung der praktischen Güte der summativen Evaluationsmethode, sowie als praktische Entscheidungshilfe dienen (siehe Bühner, 2011; Döring & Bortz, 2016; Lienert & Raatz, 1998). Unter Berücksichtigung der Anforderungen aus dem Projekt 3D-GUIde muss lediglich das Nebengütekriterium Effizienz (d.h. zeitliche Anwendungseffizienz) erfüllt sein, um einer summativen Evaluationsmethode praktische Güte aus Anwendungsprojektsicht attestieren zu können. Alle formalen summativen Gütekriterien werden nun nachfolgend im Detail vorgestellt.

3.4.1 Hauptgütekriterium Objektivität

Unter dem Begriff *Objektivität* versteht man das Ausmaß, in dem die Ergebnisse einer summativen Evaluationsmethode vom Versuchsleiter (d.h. Evaluator: Durchführender und Auswertender) oder von der Versuchssituation unabhängig sind (Bühner, 2011; Eid &

Schmidt, 2014). Zur Bestimmung der Objektivität kommen oft Maße der Beurteilerübereinstimmung zum Einsatz (Moosbrugger & Kelava, 2007). Lienert und Raatz (1998) sprechen bei Objektivität deswegen auch von interpersoneller Übereinstimmung der Evaluatoren. Als Maß der Übereinstimmung kann beispielsweise der durchschnittliche Korrelationskoeffizient zwischen den durch die verschiedenen Evaluatoren festgestellten Ergebnissen einer bestimmten Stichprobe fungieren oder es können Kappa-Statistiken genutzt werden. Bei einer objektiven Methode wird dem Evaluator demnach kein Verhaltensspielraum bei der Durchführung, Auswertung und Interpretation eingeräumt (Moosbrugger & Kelava, 2007).

Sinnvollerweise können dementsprechend drei Aspekte der Objektivität unterschieden werden mit denen man sich dem formalen Hauptgütekriterium *Objektivität* operational nähern kann (Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007). Dabei erschließt jeder dieser Aspekte einen anderen operationalen Zugang zur Bestimmung der Objektivität. Lienert und Raatz (1998) leiten daraus ab, dass die Objektivität eines Tests per se nicht existiert, sondern lediglich verschiedene damit verbundene methodische Zugänge. Es können die folgenden drei Ansätze bzw. Vorgehensweisen unterschieden werden, um die Objektivität einer Methode empirisch abschätzen zu können, wobei jedoch alle Ansätze zu unterschiedlichen Objektivitätseinschätzungen führen können (Lienert & Raatz, 1998):

Durchführungsobjektivität Dieser Ansatz beschreibt, dass eine Evaluationsmethode immer standardisiert durchgeführt und nicht von Untersuchung zu Untersuchung variieren sollte. Durchführungsobjektivität wird als das Ausmaß der Unabhängigkeit der Evaluationsergebnisse von zufälligen oder systematischen Variationen des Evaluatorverhaltens verstanden, welcher wiederum seinerseits zu Verhaltensvariationen bei den Versuchspersonen führen und somit das Ergebnis verfälschen kann (Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007). Zur Gewährleistung der Durchführungsobjektivität müssen unter anderem die Instruktion, Bearbeitungsdauer und etwaige Hilfestellungen bei der Beantwortung von Fragen identisch sein (Bühner, 2011; Eid & Schmidt, 2014).

Auswertungsobjektivität Diese Art der Objektivität ist gegeben, wenn jeder Auswerter die gleichen Punkt- oder Leistungswerte eines Probanden ermitteln kann. Es muss sich beim Durchführenden und Auswertenden nicht notwendigerweise um dieselbe Person handeln. Auswertungsobjektivität betrifft somit die numerische oder kategoriale Auswertung des beobachteten Testverhaltens nach vorgegebenen Regeln (Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007). Dies gelingt vollständig, wenn für jedes Antwortverhalten eine eindeutige Auswertungsvorschrift (z.B. Verwendung von Antwortschablonen zur Punktvergabe; automatisierte Auswertung der Bearbeitung durch das System) zur Verfügung gestellt werden kann (Bühner, 2011; Eid & Schmidt, 2014).

Interpretationsobjektivität Diese Art der Objektivität fordert, dass verschiedene Auswerter möglichst zur gleichen Beurteilung oder Interpretation der Testergebnisse kommen sollen (Eid & Schmidt, 2014). Interpretationsobjektivität betrifft somit das Ausmaß der Unabhängigkeit der Ergebnisinterpretation vom interpretierenden Evaluator. Dieser Evaluator muss nicht notwendigweise mit dem Durchführenden oder Auswertenden übereinstimmen (Lienert & Raatz, 1998). Dazu werden ausreichend große Normstichproben und ausreichend geprüfte Evaluationskriterien benötigt (Bühner,

2011; Moosbrugger & Kelava, 2007). Auf diese Weise kann jede Person gleichermaßen beurteilt werden. In der Regel wird dies durch die Normierung und gegebenenfalls durch geeignete Interpretationsbeispiele erreicht (Moosbrugger & Kelava, 2007).

3.4.2 Hauptgütekriterium Reliabilität

Unter dem Begriff *Reliabilität* versteht man das Ausmaß der Genauigkeit, mit dem eine summative Evaluationsmethode ein bestimmtes Evaluationskriterium eines Evaluationsgegenstands misst und zwar unabhängig davon, ob es sich dabei wirklich um das Kriterium handelt, dessen Erfassung beabsichtigt ist (Bühner, 2011; Eid & Schmidt, 2014; Moosbrugger & Kelava, 2007). Letzterer Aspekt wird durch das formale Hauptgütekriterium der Validität behandelt (Lienert & Raatz, 1998). Mithilfe der Reliabilität können Aussagen zur Zuverlässigkeit, Stabilität und Konsistenz der summativen Evaluationsmethode getroffen werden. Die Reliabilität eines psychologischen Tests kann dabei nicht höher ausfallen als dessen Objektivität (Lienert & Raatz, 1998). Die Reliabilität wird durch einen Reliabilitätskoeffizienten bestimmt, welcher aussagt in welchem Ausmaß unter gleichen Rahmenbedingungen gewonnene Messwerte zu ein und demselben Probanden übereinstimmen, also inwiefern das Ergebnis reproduzierbar ist (Lienert & Raatz, 1998).

Wie bei der Objektivität sollten auch bei der Reliabilität verschiedene Aspekte unterschieden werden. Dabei erschließt jeder dieser Aspekte einen anderen operationalen Zugang zur Reliabilitätsbestimmung. Lienert und Raatz (1998) leiten daraus ab, dass die Reliabilität eines Tests per se nicht existiert, sondern lediglich verschiedene damit verbundene methodische Zugänge. Es können die folgende vier Ansätze bzw. Vorgehensweisen unterschieden werden, um die Reliabilität einer Methode empirisch abschätzen zu können, wobei jedoch alle Ansätze zu unterschiedlichen Reliabilitätseinschätzungen führen können (Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007):

Retest-Reliabilität Diese Art der Reliabilität wird mithilfe der Testwiederholungsmethode bestimmt, welche vorsieht, aus einer Stichprobe von Versuchspersonen zu zwei unterschiedlichen Testzeitpunkten die selbe summative Evaluationsmethode durchführen zu lassen (d.h. unter der idealen Annahme, dass sich das durch die summative Evaluationsmethode erfasste Merkmal nicht geändert hat). Die Reliabilität wird anschließend durch Korrelation zwischen beiden Ergebnisreihen bestimmt (Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007).

Paralleltest-Reliabilität Laut Moosbrugger und Kelava (2007) können eine Reihe von reliabilitätsverändernden Einflüssen (z.B. Übungseffekte und Merkmalsveränderungen) dadurch kontrolliert werden, dass die Reliabilität mithilfe der Paralleltestreliabilität bestimmt wird. Diese Art der Reliabilität wird bestimmt, indem man bei einer Stichprobe zwei inhaltlich möglichst ähnliche summative Evaluationsmethoden anwendet und deren Ergebnisse anschließend miteinander korreliert (Bühner, 2011; Lienert & Raatz, 1998).

Testhalbierungs-Reliabilität In vielen Kontexten lässt sich ein Test oftmals nicht wiederholen oder parallele Testformen herstellen (Moosbrugger & Kelava, 2007), da Versuchspersonen an weiteren Terminen nicht zur Verfügung stehen oder die parallele Durchführung aus Kosten- und Kapazitätsgründen schwierig ist. In solchen

Fällen lässt sich die Reliabilität anhand der Testhalbierungs-Reliabilität bestimmen. Unter Testhalbierungs-Reliabilität versteht man die Korrelation zwischen zwei Ergebnishälften der gleichen Evaluationsmethode. Die Ergebnislänge wird dabei als Faktor in die Berechnung mit aufgenommen, um die verminderte Testhalbierungs-Reliabilität auf die ursprüngliche Testlänge hin korrigieren zu können. Führt man eine solche Korrektur durch, führt dies laut Moosbrugger und Kelava (2007) zu einer Überschätzung der Testhalbierungs-Reliabilität. Eine entsprechende Korrekturmöglichkeit bietet die Spearman-Brown-Formel (Moosbrugger & Kelava, 2007).

Innere Konsistenz Im Gegensatz dazu wird bei der Konsistenzanalyse das Ergebnis der Methode nicht in zwei oder mehr äquivalente Teile aufgespalten, sondern in ebenso viele Teile wie Aufgaben untergliedert, und stellt daher einer Verallgemeinerung der Testhalbierungs-Reliabilität dar (Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007). Hierbei werden die Elemente einer summativen Evaluationsmethode (z.B. Items eines Fragebogens) als multipel halbierte Ergebnisteile aufgefasst und die Reliabilität über bestimmte Zielwerte dieser Testelemente (z.B. Aufgabenschwierigkeits- und Trennschärfestatistiken) auf indirektem Wege mithilfe von Korrelationen ermittelt (Lienert & Raatz, 1998). Hierbei kommt als Maß für die interne Konsistenz häufig ein Cronbach-Alpha-Koeffizient zum Einsatz (Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007).

3.4.3 Hauptgütekriterium Validität

Die Validität gilt als das wichtigste formale Hauptgütekriterium von psychologischen Tests (Bühner, 2011; Döring & Bortz, 2016; Eid & Schmidt, 2014; Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007) und damit auch von summativen Evaluationsmethoden. Unter dem Begriff *Validität* versteht man das Ausmaß der Genauigkeit, mit dem eine Evaluationsmethode dasjenige Merkmal oder diejenige Verhaltensweise, die gemessen werden soll, tatsächlich auch misst oder vorhersagt (Bühner, 2011; Lienert & Raatz, 1998). Demzufolge setzt ein solcher psychologischer Test auch notwendigerweise eine hohe Objektivität und Reliabilität voraus, um überhaupt eine hohe Validität besitzen zu können (Moosbrugger & Kelava, 2007). Laut Lienert und Raatz (1998) entbindet somit die formale Prüfung der Validität in gewissem Maße von der Überprüfung der übrigen beiden formalen Hauptgütekriterien.

Auch bei der Validität existieren wie bei der Reliabilität und Objektivität zuvor unterschiedliche Aspekte, über die man sie bewerten kann. Dabei erschließt jeder dieser Aspekte einen anderen operationalen Zugang zur Bestimmung der Validität. Lienert und Raatz (1998) leiten daraus ab, dass die Validität eines Tests per se nicht existiert, sondern lediglich verschiedene damit verbundene methodische Zugänge. Dabei lässt sich zwischen Inhalts-, Kriteriums- und Konstruktvalidität unterscheiden, wobei alle Ansätze zu unterschiedlichen Validitätseinschätzungen führen können (Lienert & Raatz, 1998). Letztere wird jedoch als der bedeutsamste Aspekt angesehen, da sowohl Inhalts- als auch Kriteriumsvalidität darin integriert sind (Lienert & Raatz, 1998).

3.4.3.1 Inhaltsvalidität

Man spricht von Inhaltsvalidität, wenn eine Evaluationsmethode das Zielmerkmal (d.h. Evaluationskriterium) theoretisch vollständig erfassen kann (Bühner, 2011; Döring & Bortz, 2016), weswegen man hier auch oftmals von Repräsentationsschluss spricht (Moosbrugger & Kelava, 2007). Inhaltsvalidität ist dabei entweder offensichtlich (d.h. Augenscheinvalidität) oder wird durch Experten als valide beurteilt. Die Validierung erfolgt dazu entweder nach fachlichen und logischen Überlegungen (Fisseni, 1997) oder es werden Expertenratings hinzugezogen (Lienert & Raatz, 1998). Sie wird somit in der Regel nicht über einen numerischen Kennwert bzw. den Vergleich mit einem Außenkriterium erfasst, sondern mit oder ohne Einschränkungen akzeptiert oder verworfen (Bühner, 2011). Besteht Expertenkonsens bezüglich der repräsentativen Erfassung aller relevanten inhaltlichen Aspekte durch die Evaluationsmethode, wird die Passung der Evaluationsmethode zum zu erfassenden Konstrukt als hoch eingeschätzt, was eine hohe Inhaltsvalidität indiziert (Moosbrugger & Kelava, 2007).

3.4.3.2 Kriteriumsvalidität

Laut Moosbrugger und Kelava (2007) bezieht sich die Kriteriumsvalidität auf die praktische Anwendbarkeit einer summativen Methode (d.h. Test) für die Vorhersage der Ausprägung eines Merkmals (z.B. intuitive Benutzung), anders ausgedrückt auf ein Verhalten außerhalb der Versuchssituation. Eine summative Evaluationsmethode (d.h. Test) weist Kriteriumsvalidität auf, wenn aufgrund des Nutzerverhaltens innerhalb der Versuchssituation erfolgreich auf ein Nutzerverhalten außerhalb der Versuchssituation geschlossen werden kann. Letzteres fungiert dann als ein Außenkriterium (Moosbrugger & Kelava, 2007). Kriteriumsvalidität kennzeichnet somit den Zusammenhang (d.h. Korrelation) der Evaluationsergebnisse (d.h. Testwert) einer Stichprobe (z.B. Ausprägung des Evaluationskriteriums intuitiver Benutzung) mit einem oder mehreren anderer Außenkriterien, mit dem die Evaluationsmethode aufgrund ihres Messanspruchs korrelieren sollte. Dies wird auch als Korrelationsschluss bezeichnet, da die Prüfung der Kriteriumsvalidität lediglich auf Zusammenhängen zwischen dem Ergebnis der neuen untersuchten Methode (d.h. Testwert) und dem Ergebnis eines echten Außenkriteriums (d.h. Zielwert) oder den Ergebnissen mehrerer Quasi-Außenkriterien zur Approximation eines echten Kriteriums basiert (Bühner, 2011). Es können hinsichtlich des Status der für eine Kriteriumsvalidierung genutzten Außenkriterien zwischen zwei Arten der Kriteriumsvalidität unterschieden werden (Beauducel & Leue, 2014; Döring & Bortz, 2016; Lienert & Raatz, 1998):

Innere kriteriumsbezogene Validität Der Testwert der summativen Evaluationsmethode wird mit den Zielwerten von Quasi-Außenkriterien korreliert.

Äußere kriteriumsbezogene Validität Der Testwert der summativen Evaluationsmethode wird mit dem Zielwert eines echten Außenkriteriums korreliert.

Ferner sind je nach Zeitpunkt der Erfassung des Außenkriteriums drei unterschiedliche Arten der Kriteriumsvalidität zu unterscheiden (Bühner, 2011; Döring & Bortz, 2016; Lienert & Raatz, 1998):

Vorhersagevalidität (d.h. prognostische Validität) Der Zielwert des Außenkriteriums wird zeitlich nach dem Testwert der zu validierenden Methode erhoben und im Anschluss Zusammenhänge ermittelt (z.B. ein Intelligenztest wird über die später erhaltene Abiturnote validiert).

Übereinstimmungsvalidität (d.h. konkurrente Validität) Der Zielwert des Außenkriteriums wird zum gleichen Zeitpunkt wie der Testwert der zu validierenden Methode erhoben und Zusammenhänge ermittelt (z.B. Konzentrationsleistung bei einer Klausur wird anhand der aktuellen Noten validiert).

Retrospektive Validität Der Zielwert des Außenkriteriums wird zeitlich vor dem Testwert der zu validierenden Methode erhoben (z.B. Intelligenztest wird während des Studiums erhoben und mit den Schulnoten des zurückliegenden Abiturs validiert).

Außerdem kann die zu validierende Methode die Varianzaufklärung bezüglich eines Zielwerts (z.B. intuitive Benutzung) erhöhen, wenn sie mit einer anderen Methode zusammen eingesetzt wird (Bühner, 2011; Döring & Bortz, 2016):

Inkrementelle Validität Durch die zusätzliche Berücksichtigung der zu validierenden summativen Evaluationsmethode kann ein Zielwert signifikant besser vorhersagt werden, als wenn man man lediglich eine bereits etablierte summative Evaluationsmethode berücksichtigt.

3.4.3.3 Konstruktvalidität

Laut Moosbrugger und Kelava (2007) beschäftigt man sich unter dem Aspekt der Konstruktvalidität mit der theoretischen Fundierung des von einer summativen Methode (d.h. Test) tatsächlich gemessenen Merkmals (z.B. intuitive Benutzung). Im Gegensatz zur Inhaltsvalidität, für die kein numerischer Kennwert auf Basis eines Vergleichs mit einem Außenkriterium herangezogen werden kann und die daher letztendlich immer auf einer subjektiven Einschätzung basiert, kann die Konstruktvalidität und Kriteriumsvalidität durch den Abgleich mit einem Außenkriterium intersubjektiv empirisch erfasst werden (Figl, 2010). Konstruktvalidität liegt vor, wenn eine summative Evaluationsmethode das erfasst, was sie auch erfassen soll (Bühner, 2011). Zur Konstruktvalidierung leitet man basierend auf dem Konstrukt (z.B. intuitiver Benutzung) und der damit verbundenen Theorie Hypothesen ab, die unter Einsatz der zu validierenden summativen Evaluationsmethode empirisch überprüft werden. Eine summative Evaluationsmethode (d.h. Test) weist Konstruktvalidität auf, wenn der Rückschluss vom Nutzerverhalten innerhalb der Versuchssituation auf zugrunde liegende psychologische Persönlichkeitsmerkmale (d.h. Konstrukte) wissenschaftlich fundiert ist (Moosbrugger & Kelava, 2007). Sie stellt demnach eine Synthese aus Inhalts- und Kriteriumsvalidität dar (Döring & Bortz, 2016).

Zur Konstruktvalidierung können diverse methodische Wege genutzt werden. Dabei gelten *logische Analysen*, *empirisch-korrelationsstatistische* und *experimentelle Ansätze* als gleichbedeutend (Lienert & Raatz, 1998). Wird jedoch die Konstruktvalidität nur über die Kriteriumsvalidität und damit auf empirisch-korrelationsstatistische Weise definiert, besteht die Gefahr eines Zirkelschlusses, wenn nicht die theoretische Fundierung und die wissenschaftliche Güte der genutzten Außenkriterien sichergestellt werden kann (z.B. muss

die Theorie hinter der zu evaluierenden Evaluationsmethode und den Außenkriterien zusammenpassen und letztere bereits hinsichtlich der Hauptgütekriterien evaluiert worden sein). Im Gegensatz zur Kriteriumsvalidität, bei der die Frage der psychologischen Bedeutung eines Evaluationsergebnisses oder Außenkriteriums nicht im Fokus steht (d.h. praktische Anwendbarkeit des Tests für Vorhersage steht im Vordergrund und nicht theoretische Hintergründe), ist die psychologische Analyse der summativen Evaluationsmethode und damit des zugrunde liegenden Konstrukts gerade das Ziel der Konstruktvalidierung. Alle Methoden zur Konstruktvalidierung ergänzen sich, da sie das zu validierende Konstrukt von verschiedenen Seiten beleuchten. Es existiert somit kein umfassendes Maß der Konstruktvalidität. Prinzipiell lässt sich bei der Beurteilung der Konstruktvalidität zwischen *struktursuchenden* und *strukturprüfenden* Ansätzen unterscheiden (Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007).

Der erste der beiden Ansätze basiert dabei auf einer struktursuchenden deskriptiven Vorgehensweise. Üblicherweise wird dabei überprüft, ob das Ergebnis mit anderen Messergebnissen konstruktverwandter Verfahren erwartungsgemäß hoch und mit konstruktfernen Verfahren niedrig korreliert (Döring & Bortz, 2016). Dazu werden nach Campbell und Fiske (1959) die beiden Grundvoraussetzungen von Konstruktvalidität folgendermaßen formuliert:

Konkurrente bzw. konvergente Validität Der Testwert der zu validierenden summativen Evaluationsmethode soll eng mit den Zielwerten konstruktverwandter, summativer Evaluationsmethoden relativ hoch korrelieren. Der erwartete Zusammenhang sollte dabei auf jeden Fall höher als bei der Korrelation mit konstruktfremden Methoden ausfallen (Bühner, 2011; Döring & Bortz, 2016; Lienert & Raatz, 1998).

Diskriminante bzw. divergente Validität Der Testwert der zu validierenden summativen Evaluationsmethode soll mit den Zielwerten konstruktfremder, summativer Evaluationsmethoden gar nicht oder nur gering korrelieren. Der erwartete Zusammenhang sollte dabei auf jeden Fall geringer als bei der Korrelation mit konstruktnahen Methoden ausfallen (Bühner, 2011; Döring & Bortz, 2016; Lienert & Raatz, 1998).

Eine gute Konstruktvalidierung ist nur dann möglich, wenn aus dem, durch die zu validierende summative Evaluationsmethode erfassten, Zielkonstrukt (z.B. intuitive Benutzung) ein theoretisch und/oder empirisch gut fundiertes Netz an Hypothesen a priori ableitbar ist, welches Richtung und Enge der Zusammenhänge mit anderen konstruktnahen summativen Evaluationsmethoden (d.h. objektive Evaluationsmethoden, die mentale Beanspruchung erfassen; subjektive Evaluationsmethoden, die das Gefühl von Flüssigkeit erfassen) im Sinne konvergenter und divergenter Validität beschreibt (Cronbach & Meehl, 1955; Döring & Bortz, 2016; Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007). Auf diese Weise wird das Konstrukt sowie damit verbundene konstruktnahe (d.h. ähnliche) und konstruktferne (d.h. unähnliche) Methoden in ein *nomologisches Netzwerk* eingebettet (Cronbach & Meehl, 1955; Moosbrugger & Kelava, 2007). Dabei fasst ein solches Netzwerk die Gesetzmäßigkeiten zusammen, unter deren Einfluss das zu untersuchende Konstrukt steht und beinhaltet damit theoriegeleitete Erwartungen zu Ursachen und Konsequenzen eines Phänomens (Möhring & Schlütz, 2013). Auf diese Weise werden positive, negative und Nullbeziehungen zwischen der zu validierenden Methode und anderen empirischen Indikatoren beschrieben und getestet, worüber festgestellt werden kann, ob sich die prognostizierten Zusammenhänge empirisch beobachten lassen (Krohne & Hock, 2007).

Einen speziellen Ansatz dieses strukturSuchenden Vorgehens stellt die Multitrait-Multimethod-Methode dar, welche systematisch die konvergente und divergente Konstruktvalidität einer Methode untersucht (Campbell & Fiske, 1959). Diese Validierungsmethode beruht auf der Annahme, dass Testwerte einer Methode nur dann korrekt sind, wenn zum einen Messungen desselben Konstrukts mit verschiedenen Methoden zu hoher Merkmalskonvergenz (d.h. konvergente Validität) führen und zum anderen eine Unterscheidung zwischen inhaltlich unterschiedlichen Konstrukten, sowohl innerhalb jeder Methode als auch zwischen den einzelnen Methoden, beobachtet werden kann (d.h. divergente Validität) (Schermelleh-Engel & Schweizer, 2008). Man kann auf diese Weise verhindern, dass hohe Korrelationen aufgrund nicht unterschiedlicher Methoden irrtümlicherweise im Sinne einer hohen konvergenten Validität interpretiert und damit auch systematische Messfehler bzw. Methodeneffekte (d.h. Fehler gehen auf die Art der Messmethode zurück) berücksichtigt werden (Döring & Bortz, 2016; Schermelleh-Engel & Schweizer, 2008). Um die Multitrait-Multimethod-Methode anwenden zu können, müssen mindestens zwei Konstrukte (d.h. Multitrait; z.B. Kreativität, Kooperationsbereitschaft) durch mindestens zwei Methoden (d.h. Multimethod; z.B. Eigenbeurteilung durch Fragebogen, Fremdbeurteilung durch Expertenbeobachtung) gleichzeitig gemessen werden. Dabei wird jede Methode systematisch mit jedem Konstrukt kombiniert und die damit verbundenen Testwerte verglichen (Döring & Bortz, 2016; Schermelleh-Engel & Schweizer, 2008). Weitere spezielle strukturSuchende Ansätze (z.B. exploratorische Faktorenanalysen) sind detailliert bei Moosbrugger und Kelava (2007) zu finden und werden im Rahmen dieser Arbeit nicht weiter thematisiert.

Neben dem strukturSuchenden Ansatz, bei dem die Konstruktvalidität anhand der konvergenten und divergenten Validität bewertet wird, existieren auch strukturPrüfende Verfahren mit denen inferenzstatistische Schlüsse bezüglich der Konstruktvalidität gezogen werden können (Moosbrugger & Kelava, 2007). Diese Verfahren, die häufig bei Fragebögen eingesetzt werden, setzen jedoch spezielle Testmodelle voraus, die einen eindeutigen und statistisch prüfbaren Zusammenhang zwischen zuvor genau definierten, latenten Merkmalen (z.B. intuitive Benutzung) und den manifesten Itemvariablen (z.B. Test-items) herstellen. Auf deren Basis kann mit einer konfirmatorischen Faktorenanalyse die Konstruktvalidität einer Methode nachgewiesen werden (Moosbrugger & Kelava, 2007; Moosbrugger & Schermelleh-Engel, 2012). Für weitere Details bezüglich einer Faktorenanalyse wird auf Moosbrugger und Schermelleh-Engel (2012) und Field (2017) verwiesen, da im Rahmen dieser Arbeit aufgrund der inhärenten Komplexität des Konstrukts intuitiver Benutzung ein strukturSuchender Ansatz mithilfe eines nomologischen Netzes (d.h. Nachweis der konvergenten und divergenten Validität) verfolgt wird.

3.4.4 Nebengütekriterien

Neben den eben vorgestellten formalen Hauptgütekriterien, die über die wissenschaftliche Güte einer summativen Evaluationsmethode entscheiden, können im Nachgang noch formale Nebengütekriterien hinzugezogen werden, sofern die Hauptgütekriterien als erfüllt gelten (Bühner, 2011; Lienert & Raatz, 1998). Diese Nebengütekriterien dienen der Sicherstellung der praktischen Güte der summativen Evaluationsmethode und deren Auswahl kann, wie bereits angesprochen, in Abhängigkeit vom konkreten Anwenderinteresse variieren (Bühner, 2011). Im Gegensatz zu den vorgestellten Hauptgütekriterien sind die

Nebengütekriterien eher von qualitativer Natur und daher weniger leicht quantifizierbar (Döring & Bortz, 2016). Unabhängig vom Anwenderinteresse werden bei unterschiedlichen Sammlungen von Nebengütekriterien den folgenden vier Nebengütekriterien eine besonders hohe Bedeutung zugeordnet (z.B. Döring & Bortz, 2016; Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007):

Normierung Eine summative Evaluationsmethode ist normiert, wenn aktuelle Normen (d.h. durchschnittliche Evaluationsergebnisse repräsentativer Vergleichsstichproben) vorliegen, die eine Einordnung individueller Ergebniswerte erlauben (d.h. normorientiertes Testen). Der Ergebnisrohwert wird anhand der zur Verfügung stehenden Normwerte in einen standardisierten Ergebniswert (z.B. z-Wert) überführt (Döring & Bortz, 2016).

Effizienz Eine summative Evaluationsmethode ist effizient, wenn sie in Relation zum Erkenntnisgewinn (d.h. wissenschaftliche Güte) eine kurze Ausführungszeit (d.h. Zeit für Durchführung, Auswertung und Interpretation) beansprucht, wenig Material (z.B. Hard- und Software, Räume, Versuchsleiter, Versuchspersonen) benötigt, einfach zu nutzen ist, als Gruppe durchführbar und zeitlich effizient auszuwerten ist (Döring & Bortz, 2016).

Nützlichkeit Eine summative Evaluationsmethode ist nützlich, wenn sie ökonomisch ist und/oder ein für die Praxis relevantes Konstrukt (z.B. intuitive Benutzung) misst, für das bislang überhaupt keine summative Evaluationsmethode oder nur eine mit beschränkter wissenschaftlicher Güte vorlag (Döring & Bortz, 2016).

Vergleichbarkeit Eine summative Evaluationsmethode ist vergleichbar, wenn Paralleltests oder validitätsähnliche Tests eine intraindividuelle Reliabilitätskontrolle zulassen (Döring & Bortz, 2016).

Lediglich das Nebengütekriterium *Effizienz* thematisiert die zeitliche Anwendungseffizienz der Methode, wobei hier die Ausführungszeit unter Berücksichtigung der Anforderung aus dem Projekt 3D-GUIde zentral ist. Neben diesen vier grundlegenden Nebengütekriterien finden sich noch weitere Nebengütekriterien in der Forschungsliteratur, mit denen sich die allgemeine Forschungsethik bewerten lässt, sofern diese aus Anwenderinteresse einen kritischen Aspekt darstellt (Döring & Bortz, 2016). Diese folgenden Nebengütekriterien werden im Rahmen des Projekts 3D-GUIde zwar nicht explizit gefordert, sollen aber aufgrund deren universeller Relevanz immer mitbedacht werden:

Zumutbarkeit Eine summative Evaluationsmethode ist zumutbar, wenn sie die Probanden zum Erkenntnisgewinn nicht übermäßig in zeitlicher, physischer und psychischer Hinsicht belastet (Döring & Bortz, 2016).

Nicht-Verfälschbarkeit Eine summative Evaluationsmethode ist nicht-verfälschbar, wenn es Probanden nicht oder kaum gelingen kann, absichtlich ein besonders gutes oder schlechtes Evaluationsergebnis herbeizuführen, welches dem Evaluator nicht als unplausibel oder gefälscht auffällt (Döring & Bortz, 2016).

Testfairness Eine summative Evaluationsmethode ist fair, wenn sie alle Nutzergruppen, für die sie genutzt wird, nicht systematisch aufgrund ihrer ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppenzugehörigkeit diskriminiert (Döring & Bortz, 2016).

3.5 Güte von Methoden zur formativen Evaluation intuitiver Benutzung

Auf Basis aktueller Reviews innerhalb des Forschungsfeldes (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) kann laut aktuellem Kenntnisstand keine dedizierte formative Evaluationsmethode für intuitive Benutzung identifiziert werden, weswegen das Forschungsfeld zu intuitiver Benutzung gewöhnliche empirische Nutzertests mit verschiedenen Formen des Think-Aloud-Protokolls für diesen Zweck adaptiert (Blackler et al., 2018), so wie sie typischerweise zur formativen Evaluation im allgemeinen Usability-Bereich eingesetzt werden (siehe Sarodnick & Brau, 2006).

Das parallele Ausführen eines Think-Aloud-Protokolls während der Aufgabenbearbeitung (z.B. Blackler et al., 2010) wird dabei am häufigsten als formative Evaluationsmethode eingesetzt (Blackler et al., 2018). Ein retrospektives Think-Aloud-Protokoll, in welchem Nutzer basierend auf einer Videoaufzeichnung ihr Handeln im Nachgang kommentieren müssen, ist ebenfalls in Verwendung (z.B. Lawry, 2012; Reinhardt et al., 2018). Darüber hinaus wird bei den Nutzertests auch gemeinsames Explorieren (d.h. Co-Discovery) genutzt (z.B. Desai et al., 2015). Hierbei werden zwei Nutzer angehalten, eine Aufgabe gemeinsam auszuführen und dabei gleichzeitig ihre Eindrücke, Enttäuschungen und Denkprozesse durch ein paralleles Think-Aloud-Protokoll offenzulegen. Unabhängig davon welches Think-Aloud-Protokoll eingesetzt wird, ist es bei empirischen Nutzertests nicht unerheblich, wer diesen durchführt. Laut einer Reihe von Studien (z.B. Chattratichart & Brodie, 2004; Koutsabasis et al., 2007; Lindgaard, 2006) können Nutzertests im allgemeinen Usability-Bereich (d.h. nicht speziell intuitive Benutzung) zwar wissenschaftliche Güte hinsichtlich Gründlichkeit und Gültigkeit zugesprochen werden, es mangelt ihnen aber an Zuverlässigkeit (d.h. verschiedene Evaluatoren kommen nicht zu denselben Nutzungsproblemen) bei der Anwendung (z.B. Molich & Dumas, 2008; Molich, Ede, Kaasgaard, & Karyukin, 2004; Molich et al., 1999).

Der Nutzertest kam mit unterschiedlicher Instruktion (d.h. paralleles, retrospektives oder kooperatives Think-Aloud-Protokoll) zur formativen Evaluation intuitiver Benutzung unter anderem bei CUIs (z.B. Reinhardt et al., 2018), Spielsachen (z.B. Desai et al., 2015), Weckern (z.B. Lawry, 2012), Kameras (z.B. Blackler et al., 2010; Lawry, 2012), Fernbedienungen (z.B. Blackler et al., 2010), Gesundheitsprodukten (z.B. Palmer, 2018) zum Einsatz. Eine vollständige Auflistung ist bei Blackler et al. (2018) zu finden. Wie bereits in Kapitel 2 angesprochen, kann das Ausführen eines parallelen Think-Aloud-Protokolls die intuitive Benutzung des Nutzers beeinflussen und somit auch einen Einfluss auf die gefundenen Nutzungsprobleme haben (Reinhardt et al., 2018). Darüber hinaus kann eine kooperative Exploration des Systems einen Nutzer generell von der eigenen intuitiven Benutzung abhalten, wenn der andere Nutzer die Aufgabe bereits zuvor mit seiner eigenen Intuition gelöst hat. Es ist dadurch sehr schwierig, da beide Nutzer gleichermaßen ihre Denkprozesse bei der Methode parallel zur Aufgabenerledigung zum Ausdruck bringen müssen (Lim, Ward, & Benbasat, 1997).

Da dementsprechend generell von einem parallelen oder kooperativen Think-Aloud-Protokoll abzuraten ist, sollte besser auf ein retrospektives Protokoll zurückgegriffen werden, so wie es bereits von Lawry (2012) und Reinhardt et al. (2018) im Forschungsfeld der intuitiven Benutzung eingesetzt wird. Jedoch wird dieses Protokoll, wie im allgemeinen Usability-Bereich auch, aufgrund der schlechteren zeitlichen Effizienz kaum angewendet

(Blackler et al., 2018), obwohl im Usability-Bereich dessen höhere wissenschaftliche Güte gegenüber einem parallelen Think-Aloud-Protokoll bereits schon mehrfach demonstriert werden konnte (z.B. V. A. Bowers & Snyder, 1990; Hertzum, Hansen, & Andersen, 2009; Kuusela & Pallab, 2000; Van den Haak & De Jong, 2003; Van Den Haak, De Jong, & Jan Schellens, 2003; Van den Haak, de Jong, & Schellens, 2004). Da es sich beim Nutzertest mit retrospektivem Think-Aloud-Protokoll streng genommen um keine dedizierte, die Eigenheiten intuitiver Benutzung berücksichtigende, formative Evaluationsmethode handelt, wurde noch nicht dessen wissenschaftliche Güte formal im Bereich intuitiver Benutzung nachgewiesen (vergleiche Blackler, 2018; Blackler & Hurtienne, 2007; Blackler & Popovic, 2015). Da es sich jedoch intuitive Benutzung als Subkonzept von Usability verstehen lässt (Hurtienne, 2011) und damit verbundene Evaluationsmethoden dementsprechend dasselbe formative Evaluationsziel besitzen, nämlich einen möglichst großen Anteil real existierender, die Systemnutzung beeinträchtigender, Nutzungsprobleme zu finden, kann ein Nutzertest mit retrospektivem Think-Aloud-Protokoll dennoch als Quasi-Außenkriterium zur formativen Evaluation intuitiver Benutzung angesehen werden.

Wie bereits in Teilabschnitt 3.2.5 dieser Arbeit erwähnt, ist es bei der formativen Evaluation eh nahezu unmöglich ein echtes Außenkriterium zu finden, weswegen es im Forschungsfeld der Usability Usus ist, dieses anhand eines oder mehrerer Quasi-Außenkriterien zu approximieren (z.B. Blandford et al., 2008; Hartson et al., 2001). Da im Bereich intuitiver Benutzung unter Berücksichtigung aktueller Reviews (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) und einer eigenen unabhängigen Recherche keine weiteren formativen Methoden existieren, kann lediglich der Nutzertest mit retrospektivem Think-Aloud-Protokoll zur Approximation eines echten Außenkriteriums genutzt werden. Er stellt daher auch den alleinigen formativen Benchmark im Forschungsfeld zu intuitiver Benutzung dar, der zur formativen Evaluation intuitiver Benutzung von 3D-CUI-Interaktionslösungen im Rahmen des Projekts 3D-GUIde eingesetzt werden kann. Aufgrund der Tatsache, dass dieser formative Benchmark jedoch eine geringe zeitliche Anwendungseffizienz aufweist und im Projekt 3D-GUIde eine zeitliche anwendungseffiziente formative Evaluationsmethode benötigt wird, kann die Entwicklung einer neuen formativen Evaluationsmethode namens IntuiBeat-F gerechtfertigt werden. Diese Methode soll stellvertretend für den aktuellen formativen Benchmark im Projekt 3D-GUIde zur zeitlich effizienten formativen Evaluation eingesetzt werden.

3.6 Güte von Methoden zur summativen Evaluation intuitiver Benutzung

Die aktuelle Forschung zu intuitiver Benutzung unterscheidet bei den verfügbaren summativen Evaluationsmethoden für intuitive Benutzung zwischen *objektiven* und *subjektiven* Methoden (Blackler et al., 2018). Objektive Methoden erheben, einem quantitativen Wissenschaftsverständnis (siehe Döring & Bortz, 2016) folgend, eher „harte“ Daten (d.h. direkt beobachtbare Verhaltensdaten wie Fehlerrate) und versuchen somit subjektive Störeinflüsse weitgehend auszuschalten (Döring & Bortz, 2016; Hegner, 2003). Mithilfe von subjektiven Methoden werden einem qualitativen Wissenschaftsverständnis (siehe Flick, 2006) folgend eher „weiche“ Daten (d.h. Meinungen und Ansichten wie die Selbsteinschätzung intuitiver Benutzung über einen Fragebogen oder ein Interview) gewonnen, die direkt

mit der Beurteilung des Nutzers verbunden sind (Döring & Bortz, 2016; Hegner, 2003). Objektive Methoden versuchen dabei das objektive zentrale Merkmal intuitiver Benutzung, die mentale Beanspruchung, sowie das pragmatische Merkmal der Effektivität möglichst kontinuierlich während der tatsächlichen Systemnutzung zu erfassen, um auf dessen Basis die intuitive Benutzung des Systems bewerten zu können. Subjektive Methoden versuchen hingegen das subjektive zentrale Merkmal intuitiver Benutzung, das metakognitive Gefühl von Flüssigkeit, zu messen, um auf dessen Basis die intuitive Benutzung des Systems bewerten zu können.

Im folgenden Abschnitt werden nun der aktuelle Forschungsstand zur summativen Evaluation intuitiver Benutzung auf Basis subjektiver und objektiver Methoden präsentiert. Dabei soll im kommenden Teilabschnitt die wissenschaftliche Güte vorhandener subjektiver Methoden, die von den einzelnen Forschergruppen entweder selbst entwickelt oder aus anderen Forschungsbereichen übernommen wurden, bezüglich der formalen wissenschaftlichen Gütekriterien *Objektivität*, *Validität* und *Reliabilität* bewertet und daraufhin subjektive Quasi-Außenkriterien identifiziert werden. Im Anschluss werden die Limitationen vorhandener subjektiver Methoden diskutiert und begründet, warum keine dieser subjektiven Methoden aufgrund der damit verbundenen retrospektiven Evaluation intuitiver Benutzung ausschließlich für die Evaluation der 3D-CUI-Interaktionslösungen im Rahmen des Projekts 3D-GUIde als summativer Benchmark genutzt werden kann und subjektive Methoden deswegen lediglich als Quasi-Außenkriterien fungieren können. An dieser Stelle sei anzumerken, dass für die Auswahl der subjektiven Methoden die wissenschaftliche Güte und damit der empirisch nachgewiesene Zusammenhang mit dem Konstrukt *intuitive Benutzung* ausschlaggebend war. Dementsprechend wurden keine Fragebögen außerhalb des Forschungsfelds zu intuitiver Benutzung berücksichtigt, die sich ausschließlich mit der Erfassung des Gefühls von Flüssigkeit befassen (z.B. Graf et al., 2018), da deren wissenschaftliche Güte nicht explizit bezüglich intuitiver Benutzung nachgewiesen ist und sie somit auch nicht als Quasi-Außenkriterien für eine Meta-Evaluation fungieren können.

Im darauffolgenden Teilabschnitt wird die wissenschaftliche Güte vorhandener objektiver Methoden, die von den einzelnen Forschergruppen entweder selbst entwickelt oder aus anderen Forschungsbereichen übernommen wurden, bezüglich der formalen wissenschaftlichen Gütekriterien *Objektivität*, *Validität* und *Reliabilität* bewertet und daraufhin objektive Quasi-Außenkriterien identifiziert. Im Anschluss werden die Limitationen vorhandener objektiver Methoden diskutiert und ein summativer Benchmark identifiziert, der auf Basis des aktuellen Forschungsstandes für die summative Evaluation der 3D-CUI-Interaktionslösungen im Rahmen des Projekts 3D-GUIde genutzt werden kann. Aufgrund der Tatsache, dass der aktuelle summative Benchmark nicht die vom Projekt 3D-GUIde geforderte zeitliche Anwendungseffizienz erfüllt, kann unter Berücksichtigung der Limitationen aktueller subjektiver und objektiver Methoden abschließend die Entwicklung einer neuen summativen Evaluationsmethode namens IntuiBeat-S gerechtfertigt werden. Diese Methode soll stellvertretend für den aktuellen summativen Benchmark im Projekt 3D-GUIde zur zeitlich effizienten summativen Evaluation eingesetzt werden. Werden in beiden Teilabschnitten statische Ergebnisse angeführt, handelt es sich dabei immer um statistisch signifikante Ergebnisse ($p < .05$), wenn nicht anderweitig vermerkt. Wie bereits bei den subjektiven Methoden sei angemerkt, dass für die Auswahl der objektiven Methoden die wissenschaftliche Güte und damit der empirisch nachgewiesene Zusammenhang mit dem Konstrukt *intuitive Benutzung* ausschlaggebend war. Da bisher ein solcher Zusammenhang

lediglich für Hauptaufgabenleistungsmaße und Maße zur Erfassung mentaler Effizienz im Forschungsfeld zu intuitiver Benutzung sichergestellt werden konnte (siehe Blackler, 2018), werden im Folgenden Maße zur objektiven Erfassung des Gefühls von Flüssigkeit nicht als Quasi-Außenkriterien berücksichtigt. Des Weiteren leiden derartige Maße unter eine Reihe von methodischen Einschränkungen, die ihre wissenschaftliche Güte beeinträchtigen können und sie so als Quasi-Außenkriterien für eine Meta-Evaluation ungeeignet machen (z.B. reagiert der Musculus corrugator supercilii auch allgemein auf Affekt und bildet damit nicht exklusiv das Gefühl von Flüssigkeit ab oder die Reaktionszeit zur objektiven Bestimmung des Gefühls von Flüssigkeit kann durch die wahrgenommene Ästhetik des dargebotenen Stimulus verzerrt werden; siehe Reimann, Zaichkowsky, Neuhaus, Bender, & Weber, 2010; Topolinski, Likowski, Weyers, & Strack, 2009).

3.6.1 Subjektive Methoden

Im HCI-Bereich existieren zahlreiche Fragebögen zur generellen Erfassung der Usability eines Systems (z.B. System Usability Scale, siehe Bangor, Kortum, & Miller, 2008; Brooke, 1996), jedoch existieren nur wenige Fragebögen, die im Speziellen auf die Besonderheiten intuitiver Benutzung eingehen (Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018). Zur subjektiven Erfassung intuitiver Benutzung anhand des zufriedenstellenden Gefühls von Flüssigkeit kommen verschiedene Fragebögen, wie der von der QUT-Forschergruppe entwickelte *Technology Familiarity Questionnaire* (z.B. Blackler, 2006; Blackler et al., 2010; Desai et al., 2015; Gudur, Blackler, Popovic, & Mahar, 2009; Lawry, 2012; McEwan, 2017), der von der IUUI-Forschergruppe entwickelte *Questionnaire for Intuitive Use* (z.B. Hurtienne, 2011; Naumann & Hurtienne, 2010; Reinhardt et al., 2018) und der von der INTUI-Forschergruppe entwickelte *INTUI-Fragebogen* (z.B. Diefenbach & Ullrich, 2015; Tretter et al., 2018; Ullrich, 2013) zum Einsatz. Darüber hinaus werden von den drei Forschergruppen und anderen Forschern innerhalb des Forschungsfeldes zu intuitiver Benutzung auch noch der *Subjective Mental Effort Questionnaire* (z.B. Hurtienne & Blessing, 2007; Okimoto, Silva, & Miranda, 2012; Reinhardt et al., 2018; Tscharn, Latoschik, Löffler, & Hurtienne, 2017; Winkler, Baumann, Huber, Tscharn, & Hurtienne, 2016) und der *NASA Task Load Index* (z.B. Chattopadhyay & Bolchini, 2014; O'Brien, 2018) als Fragebögen zur subjektiven Selbstbeurteilung angewendet, die aus anderen Forschungsgebieten der HCI übernommen wurden.

Es soll an dieser Stelle außerdem darauf hingewiesen werden, dass noch weitere im Folgenden nicht genannte Ansätze existieren, die sich beispielsweise lediglich mit der Evaluation in einer bestimmten Domäne beschäftigen (z.B. Natural User Interfaces: Macaranas et al., 2015) oder eine Kombination von bereits existierenden Fragebögen aus anderen Forschungsgebieten darstellen (z.B. Antle et al., 2009; McAran, 2018). Für eine detaillierte Übersicht dieser Arbeiten wird auf das Review von Blackler et al. (2018) verwiesen. Die oben genannten subjektiven Methoden sollen nun detailliert vorgestellt und bezüglich ihrer wissenschaftlichen Güte bewertet und ihre Limitationen diskutiert werden. Darüber hinaus wird aufgezeigt, welche vorgestellten subjektiven Methoden sich als Quasi-Außenkriterien für eine Meta-Evaluation im Bereich intuitiver Benutzung eignen und warum subjektive Methoden nicht einen Benchmark für die summative Evaluation intuitiver Benutzung im Rahmen des Projekts 3D-GUIde darstellen können.

3.6.1.1 Technology Familiarity Questionnaire der QUT-Forschergruppe

Wie bereits erwähnt, trägt das Gefühl von Vertrautheit zu einem umfassenden metakognitiven Gefühl von Flüssigkeit bzw. einem Gefühl von Richtigkeit bei (Ackerman & Thompson, 2017). Es spiegelt sich damit auch als Zufriedenheitsaspekt in Form von hoher subjektiv wahrgenommener Vertrautheit in den vorgestellten Definitionen intuitiver Benutzung wider. Dieser Zusammenhang wurde bereits in Kapitel 2 dieser Arbeit vorgestellt. Vertrautheit resultiert als subjektives Gefühl aus der hohen objektiven Flüssigkeit bei der kognitiven Informationsverarbeitung und der damit verbundenen geringen mentalen Beanspruchung (siehe Teilabschnitt 2.2.2). Die konkrete Ausprägung der Vertrautheit wird wiederum von (1) der Dauer, (2) der Intensität und (3) der Diversität der Auseinandersetzung mit dem Produkt bestimmt (Hurtienne, Horn, & Langdon, 2010). Die Dauer beschreibt dabei den Nutzungszeitraum des Produkts. Intensität beschreibt wie oft das Produkt in diesem Zeitraum genutzt wurde. Die Diversität beschreibt die Anzahl der genutzten Funktionen oder Features bzw. die Anzahl der mit dem System gelösten Probleme (Blackler, 2018). Zusätzlich lassen sich Dauer, Intensität und Diversität auf verschiedenen Spezifikationsebenen erfassen, nämlich für (1) ein bestimmtes System, (2) Systeme des gleichen Typs und (3) Systeme verschiedener Typen (Blackler, 2018; Hurtienne et al., 2010).

Der Technology Familiarity Questionnaire (TFQ) wurde von Blackler (2006) entwickelt, um relevantes Vorwissen der Nutzer bezüglich eines bestimmten Systems (z.B. Autodesk AutoCAD 2018) beurteilen zu können und kommt hauptsächlich in den Arbeiten der QUT-Forschergruppe zum Einsatz (Blackler et al., 2018). Er berücksichtigt dabei die Intensität und Diversität des erworbenen Vorwissens auf allen drei genannten Spezifikationsebenen. Es existiert dabei kein universeller TFQ, sondern ein TFQ muss für jedes konkret zu evaluierende System neu angelegt werden, weswegen sich auch die wissenschaftliche Güte des Messinstruments als subjektives Maß für intuitive Benutzung bezüglich der formalen Gütekriterien *Objektivität*, *Validität* und *Reliabilität* nicht bestimmen lässt. Die Sicherstellung der wissenschaftlichen Güte eingesetzter TFQs erfolgte dementsprechend lediglich anhand der in Teilabschnitt 3.2.5 vorgestellten nicht formalen Gütekriterien (z.B. Blackler, 2006; Blackler et al., 2010; Desai et al., 2015; McEwan et al., 2014), insbesondere der Kriterien *Sensitivität* (d.h. „Kann das Maß Unterschiede bezüglich intuitiver Benutzung überhaupt erkennen?“) und *Diagnostizität* (d.h. „Kann das Maß die Ursachen für die Unterschiede bezüglich intuitiver Benutzung aufzeigen?“). Aufgrund der Tatsache, dass ein TFQ wegen der Art und Weise, wie dieser erstellt wird, in seiner Länge stark variieren kann, können an dieser Stelle auch keine Aussagen bezüglich dessen zeitlicher Anwendungseffizienz getroffen werden. Aktuelle Reviews des Feldes (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) konnten bisher auch keine empirischen Nachweise bezüglich der zeitlichen Anwendungseffizienz des Fragebogens im Vergleich zu anderen subjektiven Methoden liefern, weswegen dem Fragebogen auch keine hohe zeitliche Anwendungseffizienz attestiert werden kann.

Um einen TFQ zu erstellen, müssen vom Evaluator zunächst andere Systeme ausgewählt werden, die über ähnliche Features wie das zu evaluierende System verfügen. Diese fungieren dann als Items für den Fragebogen. Nutzer müssen im Anschluss sowohl diese nicht getesteten Systeme (z.B. Autodesk 123Design, Autodesk Fusion 360) als auch das eigentlich

zu evaluierende System (z.B. Autodesk AutoCAD 2018) bezüglich ihrer *Nutzungshäufigkeit* (d.h. „Wie häufig verwenden Sie die folgenden Systeme?“) und des dabei *genutzten Funktionsumfangs* (d.h. „Wenn Sie die folgenden Software-Anwendungen verwenden, wie viele Features nutzten Sie davon?“) beurteilen. Jedes Item muss bezüglich der Häufigkeitsdimension auf einer siebenstufigen Likertskala von „täglich“ (6) bis „niemals“ (0) und bezüglich der Funktionsdimension auf einer fünfstufigen Likertskala von „alle Features“ (4) bis „keines der Features“ (0) vom Nutzer bewertet werden. Beide Dimensionen werden im Anschluss dadurch ausgewertet, indem zunächst für jede Abstufung der betroffenen Likertskala (d.h. Häufigkeitsdimension, Funktionsdimension) die Anzahl der Nennungen über alle Items hinweg aufsummiert wird, und diese im Anschluss mit dem jeweiligen Wert der Abstufung (z.B. „6“ für täglich) multipliziert werden. Die Bildung eines TFQ-Scores wird abschließend durch Addition der beiden Dimensionen für jeden Nutzer vorgenommen. Der maximale Gesamtscore ist dabei abhängig von der Anzahl der gewählten Vergleichssysteme. Das theoretische Minimum liegt bei 0. Höhere TFQ-Werte weisen dabei auf ein höheres Maß an vorhandenem Vorwissen, einer höheren Vertrautheit und somit auf eine erhöhte Wahrscheinlichkeit intuitiver Benutzung hin (Blackler, 2006; Blackler et al., 2010).

3.6.1.2 Questionnaire for Intuitive Use der IUUI-Forschergruppe

Im Gegensatz zum TFQ, der versucht das Gefühl von Flüssigkeit über die wahrgenommene Vertrautheit des Nutzers zu erfassen, findet man in der Forschungsliteratur zu intuitiver Benutzung auch Fragebögen, die das metakognitive Gefühl von Flüssigkeit anhand von mehreren damit assoziierten Gefühlen erfassen. Der Questionnaire for Intuitive Use (QUESI) erfasst die in Kapitel 2 beschriebenen metakognitiven Gefühle (d.h. Gefühl des Irrtums, Gefühl der Vertrautheit, Gefühl der Kenntnis, Beurteilung des Lernens, Gefühl der Flüssigkeit), welche sich, wie bereits erwähnt, zu einem umfassenden metakognitiven Gefühl von Flüssigkeit bzw. Gefühl von Richtigkeit aggregieren lassen (Ackerman & Thompson, 2017). Ein hohes Maß an Flüssigkeit resultiert demnach aus einer geringen wahrgenommenen mentalen Beanspruchung (z.B. realisiert durch das QUESI-Item „Es gelang mir, das System ohne Nachdenken zu benutzen“), einem geringen wahrgenommenen Lernaufwand (z.B. realisiert durch das QUESI-Item „Es fiel mir von Anfang an leicht, das System zu benutzen“), einer hohen wahrgenommenen Vertrautheit (z.B. realisiert durch das QUESI-Item „Der Umgang mit dem System erschien mir vertraut“), einer hohen wahrgenommenen Zielerreichung (z.B. realisiert durch das QUESI-Item „Es gelang mir, meine Ziele so zu erreichen, wie ich es mir vorgestellt habe“) und einer geringen wahrgenommenen Fehlerrate (z.B. realisiert durch das QUESI-Item „Bei der Benutzung des Systems sind keine Probleme aufgetreten“), so wie es bereits in Kapitel 2 und insbesondere in Teilabschnitt 2.2.2 dieser Arbeit ausführlich beschrieben wurde.

Diese Kriterien der Zufriedenstellung bilden die fünf Subskalen des QUESI und werden anhand von 14 Items erfasst, die entlang mehrerer Validierungsstudien aus mehr als 30 Items ausgewählt wurden (Horn, 2008). Jedes dieser Items wird vom Nutzer anhand einer fünfstufigen Likertskala von „stimmt gar nicht“ (1) bis „stimmt völlig“ (5) bewertet. Bei allen Skalen weist ein hoher Skalenwert auf eine höhere Wahrscheinlichkeit intuitiver Benutzung hin. Die Werte der einzelnen Skalen erhält man durch Bildung eines Mittelwerts

über die jeweiligen Items der Skala. Der Gesamtscore wird im Anschluss durch Mittelwertbildung über alle Subskalen berechnet. Der Fragebogen fragt außerdem demographische Daten (z.B. Geschlecht und höchster Bildungsabschluss), Fakten zum evaluierten System (z.B. Hersteller) und Angaben zur bisherigen Häufigkeit der Systemnutzung des evaluierenden Systems und ähnlicher Systeme ab. Mit dem QUESI wird nicht nur eine Aufgabe mit einem System beurteilt, sondern die gesamte Systemnutzung (Naumann & Hurtienne, 2010).

Die Reliabilität des QUESI wurde in der Forschungsliteratur zu intuitiver Benutzung mithilfe einer Konsistenzanalyse nachgewiesen. Der QUESI weist dabei eine sehr gute innere Konsistenz mit einem Cronbachs α von über .90 für den gesamten Fragebogen auf, was eine hohe Reliabilität laut des Benchmarks von Field (2017) bedeutet. Die Reliabilitätskoeffizienten der einzelnen Subskalen sind bis auf die Subskala *wahrgenommene Zielerreichung* (Cronbachs $\alpha = .78$) ebenfalls von mittlerer bis hoher Ausprägung ($> .80$) und können somit laut Field (2017) als zufriedenstellend betrachtet werden (Horn, 2008; Naumann & Hurtienne, 2010).

Die Validität des QUESI als subjektives Maß für intuitive Benutzung wurde in Form von Konstruktvalidität des Fragebogens in verschiedenen Studien untersucht, indem der Fragebogen mit konstruktnahen und konstruktfernen Kriterien (d.h. Quasi-Außenkriterien) verglichen wurde (z.B. Horn, 2008; Reinhardt et al., 2018; M. L. Still & Still, 2018). Konvergente Korrelationskoeffizienten (z.B. Zusammenhang mit Anteil korrekt gelöster Aufgaben: $r \geq .33$; Zusammenhang mit SEA-Skala: $r \geq$ -.39) lagen bei der Studie von Horn (2008) laut J. Cohen (1988) daher im mittleren (d.h. $|r|$ um .30) bis hohen Bereich (d.h. $|r|$ um .50). Divergente Korrelationskoeffizienten (z.B. physische Effizienz) blieben weitestgehend nicht signifikant (Ausnahme: Zeitliche Effizienz bei der motorischen Handlungsaufführung, die in einem Fall mit $r = .28$ signifikant wurde). Die konvergenten Korrelationskoeffizienten (d.h. als Absolutwerte betrachtet) lagen jedoch in allen Fällen deskriptiv höher als die divergenten Korrelationskoeffizienten, was im Sinne eines struktursuchenden Vorgehens die Sicherstellung der Konstruktvalidität zusätzlich unterstützt (siehe Fiske, 1982; Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007).

Die Objektivität des QUESI wurde in der Literatur nicht explizit durch Beurteilerübereinstimmung nachgewiesen, sondern dessen Existenz stattdessen theoretisch begründet (z.B. Horn, 2008; Hurtienne, 2011; Naumann & Hurtienne, 2010). Aufgrund der Tatsache, dass die Reliabilität und Konstruktvalidität in der Literatur in mehreren Fällen empirisch formal nachgewiesen sind, stellt dieser Umstand jedoch kein Problem dar, da laut Lienert und Raatz (1998) die Prüfung der Validität in gewissem Maße von der formalen Überprüfung der übrigen wissenschaftlichen Gütekriterien entbindet.

Aktuelle Reviews des Forschungsfeldes zu intuitiver Benutzung (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) konnten außerdem keine empirischen Nachweise bezüglich der zeitlichen Anwendungseffizienz des Fragebogens im Vergleich zu anderen subjektiven Methoden liefern, weswegen dem Fragebogen keine hohe zeitliche Anwendungseffizienz attestiert werden kann.

3.6.1.3 INTUI-Fragebogen der INTUI-Forschergruppe

Neben dem QUESI, der von der IUUI-Forschergruppe entwickelt wurde (siehe Naumann & Hurtienne, 2010), ist noch ein weiterer Fragebogen im Forschungsbereich zu intuitiver Benutzung verbreitet, der wie der QUESI versucht, das Gefühl von Flüssigkeit anhand von mehreren Gefühlen zu erfassen. Der INTUI-Fragebogen erfasst dabei überwiegend die in Kapitel 2 beschriebenen Gefühle von Vertrautheit und Irrtum sowie das umfassende Gefühl von Flüssigkeit selbst. Ein starkes Gefühl von Flüssigkeit resultiert demnach aus einer gering wahrgenommenen expliziten Verbalisierungsfähigkeit (z.B. realisiert durch das INTUI-Item „Im Nachhinein fällt es mir schwer, die einzelnen Bedienschritte zu beschreiben“), einer hohen wahrgenommenen Mühelosigkeit (z.B. realisiert durch das INTUI-Item „Die Nutzung des Produkts fiel mir leicht“), einem hohen wahrgenommenen Bauchgefühl (z.B. realisiert durch das INTUI-Item „Bei der Nutzung des Produkts handelte ich unbewusst, ohne lange über die einzelnen Schritte nachzudenken“) und einem starken magischen Erleben (z.B. realisiert durch das INTUI-Item „Die Nutzung des Produkts war faszinierend“), so wie es bereits in Kapitel 2 und insbesondere in Teilabschnitt 2.2.2 dieser Arbeit beschrieben wurde.

Der INTUI besitzt 15 Items, die als siebenstufiges semantisches Differential organisiert sind. Die Komponente *Bauchgefühl* wird vom INTUI-Fragebogen mit vier Items erfasst (z.B. „Bei der Nutzung des Produkts ließ ich mich von meinem Verstand leiten“ vs. „...ließ mich von meinem Gefühl leiten“). Das Bauchgefühl versucht das subjektive Empfinden der objektiven Flüssigkeit bei der Anwendung des Vorwissens als Ursache für das Gefühl von Flüssigkeit zu erfassen. Im Gegenzug erfasst die Komponente *magisches Erleben* die gesamte Ausprägung des Gefühls von Flüssigkeit als Ergebnis des kognitiven Informationsverarbeitungsprozesses mit ebenfalls vier Items (z.B. „Die Nutzung des Produkts war unbedeutend“ vs. „...war begeisternd“). Die Komponente *Mühelosigkeit* erfasst mit fünf Items sowohl die wahrgenommene Zielerreichung (d.h. Gefühl des Irrtums), die wahrgenommene mentale Beanspruchung (d.h. Gefühl der Flüssigkeit) und die wahrgenommene Vertrautheit (d.h. Gefühl der Vertrautheit) (z.B. „Bei der Nutzung des Produkts erreichte ich mein Ziel nur mit Anstrengung“ vs. „...erreichte ich mein Ziel mit Leichtigkeit“). Die letzte Komponente *Verbalisierungsfähigkeit* erfasst mit drei Items, das Ausmaß der Verbalisierbarkeit der Systemnutzung und damit ebenfalls das Gefühl von Flüssigkeit, welches aus der objektiven Flüssigkeit bei der überwiegend unbewussten kognitiven Informationsverarbeitung resultiert und demnach mit geringer mentaler Beanspruchung assoziiert ist (z.B. „Im Nachhinein fällt es mir schwer, die einzelnen Bedienschritte zu beschreiben“ vs. „...ist es für mich kein Problem, die einzelnen Bedienschritte zu beschreiben“. Mit dem INTUI-Fragebogen wird nicht nur eine Aufgabe mit einem System beurteilt, sondern die gesamte Systemnutzung (Ullrich, 2014).

Im Gegensatz zum QUESI kann mithilfe des INTUI-Fragebogens die subjektive Komponente intuitiver Benutzung, das Gefühl von Flüssigkeit, explizit als mehrdimensionales Konstrukt erfasst werden. Da die INTUI-Forschergruppe auf die Bildung eines Gesamtwerts bei der Auswertung des Fragebogens verzichtet (siehe Diefenbach & Ullrich, 2015; Ullrich, 2013, 2014), kann berücksichtigt werden, dass mehrere metakognitive Gefühle zur Entstehung eines umfassenden Gefühls von Flüssigkeit führen und damit nicht alle Komponenten des Fragebogens immer gleichwertig sein müssen. Im Gegensatz zum QUESI

werden stattdessen die Wertungen für die vier Komponenten einzeln durch das Mitteln der entsprechenden Items berechnet und die Ausprägungen der einzelnen Komponenten in ihrer relativen Ausprägung zueinander interpretiert. Diese relativen Ausprägungen werden von der INTUI-Forschergruppe auch als INTUI-Pattern bezeichnet. Zusätzlich erhebt der Fragebogen trotzdem ein globales Intuitivitätsrating per Einzelitem (d.h. „Die Nutzung des Produkts war sehr intuitiv“ vs. „...war gar nicht intuitiv“), um eine Grobeinschätzung der intuitiven Benutzung leisten zu können. Dieses Item betrachtet nicht die affektiven (d.h. gefühlsmäßigen, subjektiven) zufriedenstellenden Konsequenzen intuitiver Benutzung, sondern ausschließlich die wahrgenommene Intuitivität des Produkts (siehe Diefenbach & Ullrich, 2015; Tretter et al., 2018; Ullrich, 2013, 2014). Um den INTUI-Fragenbogen als summatives Evaluationswerkzeug nutzen zu können, müssen INTUI-Patterns dementsprechend verglichen werden, weil der Fragebogen keinen einfachen Vergleich anhand eines Gesamtwertes vorsieht bzw. diesen nur mittels des genannten Einzelitems zur Verfügung stellt.

Die Reliabilität des INTUI-Fragebogens wurde wie beim QUESI auch mithilfe einer Konsistenzanalyse belegt. Die einzelnen Komponenten des INTUI-Fragenbogens weisen für sich eine sehr gute interne Konsistenz auf, wenn man die von Field (2017) hierfür vorgeschlagenen Benchmarks berücksichtigt (Cronbachs α: Mühelosigkeit: .96; Bauchgefühl: .85; Verbalisierungsfähigkeit: .84; Magisches Erleben: .81) (Ullrich, 2014).

Die Durchführungs- und Auswertungsobjektivität wurde zur Erfassung der Objektivität des INTUI-Fragebogens mithilfe entsprechender Instruktion und Auswerteregeln für den Fragebogen zumindest theoretisch sichergestellt (z.B. durch Standardisierung und Vermeidung sozialer Interaktion zwischen Nutzer und Evaluator). Werte bezüglich einer Beurteilerübereinstimmung liegen hier jedoch nicht vor (siehe Ullrich, 2014). Die Interpretationsobjektivität ist im Gegenzug zum QUESI als fragwürdig einzustufen, da die nicht normierte Interpretation der INTUI-Patterns sicherlich abhängig vom Vorwissen der Evaluatoren ist und die INTUI-Forschergruppe nur wenig Interpretationsbeispiele zur Verfügung stellt (siehe Diefenbach & Ullrich, 2015; Ullrich, 2013, 2014). Beim QUESI ist es hingegen klar, dass geringe Werte immer auch für eine geringe intuitive Benutzung des Systems durch den Nutzer sprechen.

Für den Nachweis der Validität des INTUI-Fragebogens als subjektives Maß für intuitive Benutzung wurde ebenfalls eine Konstruktvalidierung durchgeführt, nur hierfür im Gegensatz zum QUESI eine Faktorenanalyse verwendet und damit die Konstruktvalidität über die faktorielle Validität adressiert (Ullrich, 2014). Mithilfe verschiedener Rechenverfahren, die bei Field (2017) genauer beschrieben sind, wird bei einer Faktorenanalyse untersucht, welche Items eines Fragebogens das gleiche Konstrukt erfassen (Moosbrugger & Kelava, 2007). Diese Cluster von Items werden auch als Faktoren bezeichnet. Ullrich (2014) nutzte die Faktorenanalyse nicht zur explorativ zur Bestimmung seiner vier Fragebogenkomponenten *Mühelosigkeit*, *Verbalisierbarkeit*, *Bauchgefühl* und *magisches Erleben* (d.h. explorative Faktorenanalyse), sondern auch um seinen Fragebogen zu validieren (d.h. konfirmatorische Faktorenanalyse). Er führte auf dem Ergebnis von zwei Studien unabhängig voneinander konfirmatorische Faktorenanalysen durch und konnte in beiden Studien die gleichen Faktoren mit entsprechend hohen Faktorladungen extrahieren (d.h. alle Faktorladungen $\geq .7$, keine starken Querladungen). Ullrich (2014) attestierte dem

INTUI-Fragebogen aufgrund dieser konfirmatorischen Faktorenanalysen Konstruktvalidität (siehe Ullrich, 2014). Laut Tabachnick, Fidell und Ullman (2007) befinden sich die Faktorladungen von Ullrich (2014) alle im sehr guten Bereich (d.h. $> .63$).

Aktuelle Reviews des Forschungsfeldes zu intuitiver Benutzung (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) konnten außerdem keine empirischen Nachweise bezüglich der zeitlichen Anwendungseffizienz des Fragebogens im Vergleich zu anderen subjektiven Methoden liefern, weswegen dem Fragebogen keine hohe zeitliche Anwendungseffizienz attestiert werden kann.

3.6.1.4 Subjective Mental Effort Questionnaire

Das Gefühl von Flüssigkeit kann nicht nur aus verschiedenen damit verbundenen metakognitiven Gefühlen abgeleitet werden, sondern lässt sich auch direkt an der subjektiv wahrgenommenen mentalen Beanspruchung, die aus der objektiven Flüssigkeit resultiert (Ackerman & Thompson, 2017), abschätzen, so wie es bereits im Kapitel 2 ausführlich erläutert wurde. Aufgrund der Tatsache, dass sich intuitive Benutzung objektiv allein mithilfe eines Merkmals, der mentalen Beanspruchung abbilden lässt (Evans & Stanovich, 2013; Stanovich et al., 2014), wenn zusätzlich eine effektive Nutzung vorliegt (d.h. pragmatisches Merkmal intuitiver Benutzung, siehe Teilabschnitt 2.2.2), sind Fragebögen, die die subjektiven Auswirkungen dieses Merkmals erfassen, auch für die subjektiv summative Evaluation intuitiver Benutzung naheliegend (siehe Blackler et al., 2018; Hurtienne, 2011; Reinhardt et al., 2018).

Die wissenschaftliche (siehe Matthews et al., 2015) und die praktische Güte (siehe Vidulich & Tsang, 2015) von solchen Fragebögen zur Erfassung mentaler Beanspruchung wurde im HCI-Bereich unabhängig von der Forschung zu intuitiver Benutzung bereits mehrfach empirisch nachgewiesen. Solche Fragebögen kommen daher in diversen Domänen häufig zum Einsatz (Charles & Nixon, 2019; Matthews et al., 2015; Young et al., 2015). Im Forschungsbereich intuitiver Benutzung wird von allen Fragebögen, die die mentale Beanspruchung des Nutzers abbilden, der Subjective Mental Effort Questionnaire (SMEQ) (z.B. Horn, 2008; Kraft & Hurtienne, 2017; Naumann & Hurtienne, 2010; Reinhardt et al., 2018), der auch oft als Rating Scale Mental Effort (RSME) bezeichnet wird (Rubio, Díaz, Martín, & Puente, 2004; Zijlstra, 1993; Zijlstra & Van Doorn, 1985), am häufigsten eingesetzt. Auch außerhalb der Forschung zu intuitiver Benutzung ist dieser Fragebogen in der HCI auch sehr weit verbreitet (z.B. Cain, 2007; F. Chen et al., 2016; Galy et al., 2018; Rubio et al., 2004; Sauro & Lewis, 2009; Young et al., 2015). Die deutsche Übersetzung des SMEQ trägt den Namen SEA-Skala, was für „Skala zur Erfassung subjektiv erlebter Anstrengung“ steht (Eilers, Nachreiner, & Hänecke, 1986).

Die SEA-Skala besteht aus einer einzigen Skala mit neun Beschriftungen von „kaum anstrengend“ bis „außerordentlich anstrengend“ und Werten von 0 bis 220. Der Nutzer macht zur Beurteilung seiner subjektiven mentalen Beanspruchung nach einer durchgeführten Aufgabe (d.h. Handlung) ein Kreuz auf der vertikalen Linie an der entsprechenden Stelle. Im Gegensatz zur englischen Version der Skala, wo Anwender ihre eigene mentale Beanspruchung zwischen 0 und 150 beurteilen können, weist die deutsche Version eine Streckung der Skala vorwiegend in den mittleren und oberen Bereichen auf. Auf diese Weise soll eine

Unter- bzw. Überschätzung an den Endpunkten ausgeglichen werden (Eilers et al., 1986). Weitere bekannte Maße zur Beurteilung der mentalen Beanspruchung stellen die *Cooper-Harper Scale* (G. E. Cooper & Harper, 1969), die *modifizierte Cooper-Harper Scale* bzw. *Bedford Scale* (Roscoe, 1987), sowie die *POP (Prediction of Operator Performance)-Skala* (Farmer et al., 1995) dar, die aber alle im Bereich intuitiver Benutzung bisher nicht zum Einsatz kamen (siehe Blackler et al., 2018). Für eine generelle Übersicht aller in der HCI relevanten Fragebögen zur subjektiven Beurteilung mentaler Beanspruchung wird an dieser Stelle auf Cain (2007), F. Chen et al. (2016) und Young et al. (2015) verwiesen.

Die Reliabilität des SMEQ wurde außerhalb der Forschung zu intuitiver Benutzung mithilfe der Retest-Reliabilität (d.h. Testwiederholungsmethode) nachgewiesen. Die Retest-Reliabilität des SMEQ lag bei der Meta-Evaluation von Zijlstra (1993) im mittleren Bereich (r = .78), wenn man für die Interpretation die von Osburn (2000) angeführten Benchmarks für Reliabilität berücksichtigt. Die Reliabilität der deutschen Version des SMEQ, der SEA-Skala, wurde mithilfe einer Regressionsanalyse in Anlehnung an Lodge (1981) nachgewiesen (d.h. alle ermittelten Regressionskoeffizienten lagen in einem 95%-Vertrauensbereich), weswegen Eilers et al. (1986) dem Fragebogen eine angemessene Reliabilität zusprachen.

Die Validität der SEA-Skala als subjektives Maß für mentale Beanspruchung wurde in Form von Kriteriumsvalidität außerhalb der Forschung zu intuitiver Benutzung bereits in Zusammenhang mit dem NASA-TLX (d.h. Gesamtscore NASA-TLX: r = .78; Subskala *Anstrengung* des NASA-TLX: r = .81) untersucht (Seifert, 2002). Alle Korrelationskoeffizienten lagen dabei im hohen Bereich (d.h. $|r|$ um .50). Bezüglich des SMEQ, also der englischen Originalversion der Skala, liefert Zijlstra (1993) ebenfalls einen empirischen Nachweis bezüglich der Kriteriumsvalidität, indem er den Fragebogen mit einem weiteren subjektiven Fragebogen (d.h. Scale Experienced Load, r = .55) von Meijman (1991) und einem Hauptaufgabenleistungsmaß (d.h. Effektivität: Passierabstand von Schiffen, r = -.57) korrelierte. Die Korrelationskoeffizienten lagen damit laut J. Cohen (1988) ebenfalls im hohen Bereich (d.h. $|r|$ um .50). Die Kriteriumsvalidität der SEA-Skala wurde auch im Forschungsbereich intuitiver Benutzung als innere kriteriumsbezogene Validität durch eine Korrelation mit dem QUESI demonstriert (d.h. Wahrgenommene mentale Beanspruchung: r = -.32; Wahrgenommene Zielerreichung: r = -.35; Wahrgenommene Fehlerrate: r = -.39; Wahrgenommener Lernaufwand: r = -.40; Wahrgenommene Vertrautheit: r = -.39) (Horn, 2008). Diese Korrelationskoeffizienten sind laut J. Cohen (1988) als mittel (d.h. $|r|$ um .30) einzustufen. Eine Konstruktvalidierung wurde unter Berücksichtigung aktueller Reviews des Forschungsbereichs (siehe Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) noch nicht durchgeführt, ebenso wenig wie eine Kriteriumsvalidierung des SMEQ in seiner englischen Version.

Die Objektivität des SMEQ (d.h. englische Version) und der SEA-Skala (d.h. deutsche Version) wurde in der Literatur nicht explizit durch Beurteilerübereinstimmung nachgewiesen (d.h. innerhalb und außerhalb der Forschung zu intuitiver Benutzung), sondern das Vorhandensein der Objektivität der Fragebögen lediglich theoretisch begründet (z.B. Eilers et al., 1986; Zijlstra, 1993; Zijlstra & Van Doorn, 1985).

Aktuelle Reviews des Forschungsfeldes zu intuitiver Benutzung (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) konnten außerdem keine empirischen Nachweise bezüglich der zeitlichen Anwendungseffizienz des Fragebogens (d.h. deutsche

und englische Version) im Vergleich zu anderen subjektiven Methoden liefern, weswegen dem Fragebogen keine hohe zeitliche Anwendungseffizienz attestiert werden kann.

3.6.1.5 Nasa Task Load Index

Im Gegensatz zum SMEQ bzw. der SEA-Skala erfasst der NASA Task Load Index (NASA-TLX) nicht nur die subjektive mentale Beanspruchung bei der Systemnutzung, sondern betrachtet zusätzlich auch noch die damit verbundenen objektiven Ursachen. Wie bereits in Teilabschnitt 2.1.1 dieser Arbeit beschrieben, wird mentale Beanspruchung in der damit verbundenen HCI-Literatur üblicherweise dichotom aus einer eher objektiven Anforderungsperspektive (d.h. objektive Anforderungen durch die Aufgabe und Umgebung) und einer eher subjektiven Wirkungsperspektive (d.h. subjektiv wahrgenommene Auswirkung bzw. Konsequenzen der Belastung auf den Nutzer) betrachtet (Young et al., 2015).

In der HCI werden am häufigsten die *Subjective Workload Assessment Technique (SWAT)* (Reid & Nygren, 1988), das *Workload Profile (WP)* (Tsang & Velazquez, 1996) und der *NASA-TLX* (S. G. Hart & Staveland, 1988) zur Evaluation von mentaler Beanspruchung unter Berücksichtigung dieser Zwei-Komponenten-Perspektive eingesetzt (siehe Cain, 2007; F. Chen et al., 2016; Moustafa, Luz, & Longo, 2017; Rubio et al., 2004; Young et al., 2015), wovon lediglich der NASA-TLX bisher mit der Evaluation intuitiver Benutzung in Verbindung gebracht wurde (z.B. Blackler et al., 2018; Chattopadhyay & Bolchini, 2014; Hurtienne & Blessing, 2007; O'Brien, 2018; Reinhardt et al., 2018). Unabhängig von der Forschung zu intuitiver Benutzung gilt der NASA-TLX in der HCI aufgrund seiner mehrfach empirisch nachgewiesenen wissenschaftlichen Güte zur Erfassung mentaler Beanspruchung und seines hohen Verbreitungsgrads als der Goldstandard, wenn es um die Evaluation mentaler Beanspruchung auf der Basis von subjektiver Selbstbeurteilung geht (siehe F. Chen et al., 2016; Galy et al., 2012; Galy et al., 2018; Hill et al., 1992; Moustafa et al., 2017; Rizzo, Dondio, Delany, & Longo, 2016; Rubio et al., 2004; Young et al., 2015).

Der NASA-TLX nutzt sechs Dimensionen zur Beurteilung der mentalen Beanspruchung aus einer Zwei-Komponenten-Perspektive. Die Anforderungskomponente wird dabei durch die Dimensionen *geistige Anforderung* (z.B. „...war die Aufgabe leicht oder anspruchsvoll...?“), *körperliche Anforderung* (z.B. „...war die Aufgabe leicht oder schwer...?“) und *zeitliche Anforderung* (z.B. „...war die Aufgabe langsam ... oder schnell...?“) erfasst. Die Wirkungskomponente wird durch die Dimensionen *Leistung* (z.B. „Wie erfolgreich haben Sie Ihrer Meinung nach die vom Versuchsleiter (oder Ihnen selbst) gesetzten Ziele erreicht?...“), *Anstrengung* (d.h. „Wie hart mussten Sie arbeiten, um Ihren Grad an Aufgabenerfüllung zu erreichen?“ und *Frustration* (z.B. „Wie ... gestresst versus ... entspannt .. fühlten Sie sich während der Aufgabe?“) erfasst (Galy et al., 2018; S. G. Hart, 2006; S. G. Hart & Staveland, 1988; Pfendler, 1990, 1991). Jede Dimension des NASA-TLX wird vom Nutzer anhand einer 20-stufigen bipolaren Skala beurteilt und dabei ein Score von 0 (d.h. keine mentale Beanspruchung) bis 100 (d.h. hohe mentale Beanspruchung) für jede Dimension vergeben.

Wegen der vom NASA-TLX eingenommenen Zwei-Komponenten-Perspektive muss nach dem Rating der einzelnen sechs Dimensionen noch mithilfe einer Gewichtungsprozedur

herausgefunden werden, in welchem Ausmaß die verschiedenen Dimensionen in die gesamte mentale Beanspruchung einfließen und damit ein Gesamtwert ermittelt werden. Die Gewichtungsprozedur erfordert eine Reihe von paarweisen Vergleichen über alle der sechs Dimensionen, welche vor der eigentlichen Bewertung einer Dimension durchgeführt werden müssen. Anhand dieser paarweisen Vergleiche kann der Nutzer feststellen, welche der Dimensionen bezüglich seiner erlebten mentalen Beanspruchung relevanter sind als die anderen (S. G. Hart & Staveland, 1988; Rubio et al., 2004). Je häufiger eine Dimension bei den paarweisen Vergleichen gewählt wird, umso bedeutender ist sie für die Beurteilung der mentalen Beanspruchung des Nutzers und bekommt entsprechend ein höheres Gewicht zugeordnet. Jede vom Nutzer durchgeführte Aufgabe kann auf diese Weise mit einem Score von 0 bis 100 bewertet werden, indem die Wertungen der einzelnen Dimensionen mit ihren Gewichtungen multipliziert, im Anschluss aufsummiert und abschließend durch 15 (d.h. die Gesamtanzahl an paarweisen Vergleichen) geteilt werden (S. G. Hart & Staveland, 1988).

Der Vorteil einer Zwei-Komponenten-Perspektive und der von S. G. Hart und Staveland (1988) vorgeschlagenen Gewichtung der einzelnen Dimensionen liegt in der hohen Diagnostizität des Messinstruments (Dey & Mann, 2010; O'Donnell & Eggemeier, 1986; Rubio et al., 2004; Wierwille & Eggemeier, 1993). Es lässt sich damit nicht nur das Ausmaß an mentaler Beanspruchung feststellen, sondern auch der Belastungsaspekt in der Mensch-Maschine-Interaktion identifiziert werden (d.h. Anforderungskomponente), der für die erhöhte mentale Beanspruchung verantwortlich ist (Kerkau, 2006). Resultiert beispielsweise die erhöhte mentale Beanspruchung nicht nur aus der Komplexität der Aufgabe (d.h. geistige Anforderung), sondern auch aus dem Zeitdruck bei der Erledigung der Aufgabe, aufgrund dessen der Nutzer nicht genügend Zeit zum Nachdenken hatte (d.h. zeitliche Anforderung), sollte sich das sowohl in der kognitiven als auch in der zeitlichen Dimension des NASA-TLX widerspiegeln. Da die beide Anforderungen bei der Erledigung einer Aufgabe jedoch nur ins Gewicht fallen, wenn diese auch mit der Leistung des Nutzers interferieren, wird für eine Beurteilung der Gesamtarbeitsbeanspruchung neben dem Rating an sich noch eine subjektive Einschätzung der Wichtigkeit der einzelnen Dimensionen für die aktuelle Aufgabe vom Nutzer benötigt, was der NASA-TLX durch die bereits angesprochene Gewichtungsprozedur leistet. Auf diese Weise lässt sich bestimmen aus welchen Dimensionen sich die gesamte mentale Beanspruchung ableitet (S. G. Hart & Staveland, 1988).

Jedoch kritisierten eine Reihe von Wissenschaftlern diese Gewichtung hinsichtlich verschiedener Gesichtspunkte. Aufgrund der durch die 15 paarweisen Vergleiche schlechten zeitlichen Anwendungseffizienz schlug Byers (1989) eine Variante genannt *NASA-Raw Task Load Index* (NASA-RTLX) vor, bei der der Nutzer keine individuellen Gewichtungen durch Paarvergleiche vornehmen muss. Es wird stattdessen für jede Dimension die gleiche Gewichtung angenommen. Byers (1989) konnte über fünf Studien hinweg eine sehr starke Korrelation (r ~.96) zwischen der von ihm vorgeschlagenen Alternative und der Originalversion des NASA-TLX auf allen sechs Dimensionen feststellen. Andere Kritiker der Gewichtungsprozedur wie Pfendler (1991) berichteten, dass sie mithilfe eines ungerichteten Gesamtwerts durch Mittelung der sechs Dimensionen (d.h. NASA-RTLX) sowohl eine höhere Sensitivität, als auch eine höhere Reliabilität als mit dem gewichteten Gesamtwert erzielt haben. Darüber hinaus konnte von Pfendler (1991) auch keine Zunahme der Varianz zwischen den Versuchspersonen festgestellt werden. Die Korrelation zwischen den

sechs Dimensionen des NASA-TLX (d.h. gewichtet) und des NASA-RTLX (d.h. nicht gewichtet) waren wie bei Byers (1989) bei Pfendler (1990) ebenfalls sehr stark (r ~.94). Laut S. G. Hart (2006) wurden der NASA-TLX und der NASA-RTLX in 29 Studien miteinander verglichen und dabei festgestellt, dass der NASA-RTLX stärker sensitiv (z.B. Hendy, Hamilton, & Landry, 1993), genauso sensitiv (z.B. Byers, 1989) oder weniger sensitiv (z.B. Y. Liu & Wickens, 1994) als der gewichtete NASA-TLX ist. Es lässt sich daraus folgern, dass das Verhältnis der Dimensionen in der HCI bei der mentalen Beanspruchung zwar überwiegend gleich sind, aber eine Generalisierung zu Gunsten einer höheren zeitlichen Anwendungseffizienz auch zu Fehlern führen kann.

Der NASA-TLX kommt zur subjektiven Evaluation intuitiver Benutzung üblicherweise in seiner nicht gewichteten Version zum Einsatz (d.h. NASA-RTLX), welche die mentale Beanspruchung anhand eines Gesamtwerts quantifiziert (z.B. Chattopadhyay & Bolchini, 2014; O'Brien, 2018), so wie es allgemein im Forschungsfeld zu mentaler Beanspruchung Usus ist (siehe Grier, 2015; S. G. Hart, 2006). Die hohe Korrelation der gewichteten mit der nicht gewichteten Version (siehe Byers, 1989), könnte auch erklären warum zur Operationalisierung des NASA-TLX auch in Anwendungsdomänen wie Desktopsystemen auf einen globalen Gesamtwert üblicherweise zurückgegriffen wird (siehe S. G. Hart, 2006), obwohl dort eigentlich die physische Beanspruchung vernachlässigt werden könnte und sollte. Im weiteren Verlauf dieser Arbeit ist deswegen, wenn vom NASA-TLX gesprochen wird, immer seine Raw-Version gemeint, die mithilfe eines globalen Gesamtwerts operationalisiert wird. Laut meines Wissens (d.h. basierend auf aktuellen Reviews wie z. B. Blackler et al., 2018) existiert aktuell noch keine Studie, welche explizit die wissenschaftliche Güte des NASA-TLX als Maß für intuitive Benutzung zeigt.

Die Reliabilität des NASA-TLX wurde in der ursprünglichen Studie von S. G. Hart und Staveland (1988) in Form der Retest-Reliabilität empirisch nachgewiesen ($r = .83$), die unter Berücksichtigung der Benchmarks von Osburn (2000) als gut einzustufen ist. Aufgrund der hohen Korrelation zwischen dem NASA-TLX und dem NASA-RTLX kann von einer ebenfalls hohen Reliabilität der nicht gewichteten Version ausgegangen werden, wobei die innerhalb des Reviews von S. G. Hart (2006) aufgeführten Meta-Evaluationen des NASA-RTLX (z.B. Byers, 1989; Hendy et al., 1993) dies nicht mehr explizit untersucht haben.

Die Validität des NASA-TLX als subjektives Maß für mentale Beanspruchung konnte innerhalb der Human Factors Forschung zu mentaler Beanspruchung in Form von Konstruktvalidität anhand von einigen empirischen Nachweisen zu faktorieller Validität durch konfirmatorische Faktorenanalysen (d.h. mittlere Faktorladungen $\geq .9$ zur Messung mentaler Beanspruchung, keine starken Querladungen) belegt werden (z.B. Dey & Mann, 2010; Hill et al., 1992). An dieser Stelle muss jedoch angemerkt werden, dass einige Studien die Konstruktvalidität (d.h. Fragebogen misst nicht die wahrgenommene mentale Beanspruchung, sondern beispielsweise lediglich die wahrgenommene Schwierigkeit) des NASA-TLX anzweifeln (z.B. McKendrick & Cherry, 2018) oder seine hohe Verbreitung und seinen Goldstandard-Status nicht auf seine wissenschaftliche Güte, sondern lediglich auf einen Matthäus-Effekt zurückführen (z.B. De Winter, 2014). Die mittleren Faktorladungen des Fragebogens können jedoch laut Tabachnick et al. (2007) als exzellent (d.h. $> .71$) bezeichnet werden. Andere Arbeiten kritisieren die Bildung eines Gesamtwerts zur

Quantifizierung mentaler Beanspruchung (z.B. Galy et al., 2012; Galy et al., 2018) und fordern eine unabhängige Interpretation der einzelnen Dimensionen.

Die Objektivität des NASA-TLX wurde in der Literatur (z.B. Byers, 1989; S. G. Hart, 2006; S. G. Hart & Staveland, 1988) lediglich theoretisch aufgegriffen, was nicht verwunderlich ist, da ja Reliabilität und Konstruktvalidität des NASA-TLX in der Literatur überwiegend nachgewiesen sind (siehe S. G. Hart, 2006). Obwohl seine wissenschaftliche Güte im Bereich intuitiver Benutzung unter Berücksichtigung aktueller Reviews des Forschungsbereichs (siehe Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) noch nicht sichergestellt wurde und auch im HCI-Bereich die wissenschaftliche Güte des NASA-TLX von einigen Arbeiten kritisiert wird, kommt er trotzdem aufgrund seines Status als Goldstandard in der HCI (siehe Grier, 2015; S. G. Hart, 2006; Young et al., 2015) und der dort vorherrschenden schnellen Übergeneralisierung von Forschungsmethoden (siehe Teilabschnitt 3.2.5) als NASA-RTLX in Form eines Gesamturteils für die Evaluation intuitiver Benutzung in verschiedenen Bereichen zum Einsatz (siehe Blackler et al., 2018). Dabei wird der NASA-TLX wie der SMEQ üblicherweise mit anderen Maßen intuitiver Benutzung kombiniert und nicht alleine eingesetzt.

Aktuelle Reviews des Forschungsfeldes zu intuitiver Benutzung (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) konnten außerdem bisher keine empirischen Nachweise bezüglich der zeitlichen Anwendungseffizienz des Fragebogens im Vergleich zu anderen subjektiven Methoden liefern, weswegen dem Fragebogen keine hohe zeitliche Anwendungseffizienz attestiert werden kann.

3.6.1.6 Limitationen von subjektiven Methoden

Wie zu Beginn dieses Abschnitts und ausführlich in Kapitel 2 dieser Arbeit beschrieben, kann intuitive Benutzung anhand des Gefühls von Flüssigkeit subjektiv bewertet werden. Tabelle 3.2 fasst die Ergebnisse der in diesem Abschnitt vorgestellten subjektiven Methoden bezüglich der wissenschaftlichen Güte (d.h. anhand der wissenschaftlichen formalen Gütekriterien *Objektivität*, *Validität* und *Reliabilität*) und zeitlichen Anwendungseffizienz zusammen. An dieser Stelle sei angemerkt, dass innerhalb der Tabelle 3.2 ausschließlich empirische Nachweise der Validität zusammengefasst werden, die direkt im Forschungsfeld zu intuitiver Benutzung erfolgten und nicht nur im Forschungsbereich zu mentaler Beanspruchung im Bereich der HCI allgemein erbracht wurden. Bezüglich Objektivität und Reliabilität werden in der Tabelle auch empirische Nachweise zusammengefasst, die aus der allgemeinen HCI-Forschung zur mentalen Beanspruchung stammen, wenn die subjektive Methode aus diesem Feld für die summative Evaluation intuitiver Benutzung übernommen wurde. Des Weiteren konnte allen identifizierten subjektiven Methoden in empirischer Hinsicht keine hohe zeitliche Anwendungseffizienz im Vergleich zu anderen subjektiven Maßen attestiert werden, da in der Literatur entsprechende empirische Belege dafür fehlen.

Aufgrund der Tatsache, dass kein universeller TFQ erstellt werden kann, liegen auch keine empirischen Nachweise bezüglich dessen wissenschaftlicher Güte und zeitlicher Anwendungseffizienz auf Basis aktueller Reviews (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) vor (siehe Tabelle 3.2). Für den QUESI (z.B. Horn, 2008; Naumann & Hurtienne, 2010; Reinhardt et al., 2018) und den INTUI-Fragebogen

(z.B. Diefenbach & Ullrich, 2015; Ullrich, 2013, 2014) liegen im Gegenzug bereits mehrere empirische Nachweise bezüglich ihrer wissenschaftlichen Güte vor (siehe Tabelle 3.2) und beide kamen schon für die Evaluation intuitiver Benutzung bei CUIs zum Einsatz (z.B. Reinhardt et al., 2018; Ullrich, 2014), was sie beide für die Evaluation von 3D-CUI-Interaktionslösungen im Rahmen des 3D-GUIde Projekts relevant macht. Im Gegensatz zum QUESI verzichtet man beim INTUI-Fragebogen jedoch auf die Bildung eines Gesamtwertes. Dieser sieht stattdessen eine anspruchsvolle Interpretation der INTUI-Patterns vor, welche viel Expertenwissen voraussetzt. Da die verschiedenen Gefühle, die sich zu einem umfassenden Gefühl von Flüssigkeit aggregieren, abhängig vom Nutzungskontext in unterschiedlichen Anteilen zu diesem Gefühl beitragen (siehe Teilabschnitt 2.2.2), kann eine derartige Interpretation aber sinnvoll sein. Empirische Nachweise bezüglich der zeitlichen Anwendungseffizienz liegen für beide Fragebögen auf Basis aktueller Reviews (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) und der in den entsprechenden Teilabschnitten genannten Arbeiten nicht vor.

Tabelle 3.2. *Wissenschaftliche Güte bezüglich der formalen Gütekriterien Objektivität, Reliabilität und Validität, sowie bezüglich der zeitlichen Anwendungseffizienz subjektiver Evaluationsmethoden für intuitive Benutzung (anhand des subjektiven Gefühls von Flüssigkeit).*

Maß	Objektivität	Reliabilität	Validität	Zeitliche Anwendungs-effizienz
TFQ	✗	✗	✗	✗
QUESI	✗	Konsistenzanalyse (Horn, 2008)	Konstruktvalidität anhand konvergenter und divergenter Validität (Horn, 2008; Reinhardt, Kuge, & Hurtienne, 2018; M. L. Still & Still, 2018)	✗
INTUI	✗	Konsistenzanalyse (Ullrich, 2014)	Konstruktvalidität anhand konfirmatorischer Faktorenanalyse (Ullrich, 2014)	✗
SMEQ/SEA	✗	SMEQ: Retest-Reliabilität (Zijlstra, 1993) SEA: Regressionanalyse (Eilers, Nachreiner & Hänecke, 1986)	SEA: Konstruktvalidität anhand konvergenter und divergenter Validität (Horn, 2008)	✗
NASA-TLX	✗	Retest-Reliabilität (S. G. Hart & Staveland, 1988)	✗	✗

Für den SMEQ bzw. die SEA-Skala liegen sowohl innerhalb des Forschungsbereichs zu intuitiver Benutzung (z.B. Horn, 2008) als auch außerhalb (z.B. Eilers et al., 1986; Zijlstra, 1993; Zijlstra & Van Doorn, 1985) empirische Nachweise zur wissenschaftlichen Güte des Fragebogens vor (siehe Tabelle 3.2). Bezüglich der wissenschaftlichen Güte des NASA-TLX liegen lediglich außerhalb des Forschungsfeldes zu intuitiver Benutzung (z.B. Cain, 2007; F. Chen et al., 2016; Rubio et al., 2004; Young et al., 2015) mehrere empirische Belege

vor, nicht aber speziell im Forschungsfeld zu intuitiver Benutzung (siehe Tabelle 3.2). Er gilt jedoch in der HCI als der Goldstandard zur Messung mentaler Beanspruchung (siehe Grier, 2015; Young et al., 2015). Darüber hinaus ist er als subjektives Maß, das mentale Beanspruchung aus beiden Perspektiven (d.h. Anforderungs- und Wirkungsperspektive) betrachtet, laut Veltman und Gaillard (1996) in der Lage, ebenso gut mentale Beanspruchung wissenschaftlich valide zu erfassen wie Maße, die die Betrachtung lediglich aus der Wirkungsperspektive vornehmen (z.B. SMEQ).

Mithilfe des SMEQ und des NASA-TLX lässt sich die intuitive Benutzung im Gegensatz zum QUESI und dem INTUI-Fragebogen nicht nur bezogen auf das Gesamtsystem beurteilen, sondern auch auf Aufgabenebene. Beide Fragebögen kamen außerdem bereits zur Evaluation intuitiver Benutzung von CUIs zum Einsatz, was sie für die summative Evaluation von 3D-CUI-Interaktionslösungen im Rahmen des 3D-GUIde Projekts relevant und nutzbar macht (z.B. Fischbach, Neff, Pelzer, Lugrin, & Latoschik, 2013; Reinhardt et al., 2018). Empirische Nachweise bezüglich der zeitlichen Anwendungseffizienz liegen für beide Fragebögen auf Basis aktueller Reviews (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) und der in den entsprechenden Teilabschnitten genannten Arbeiten im Forschungsfeld zu intuitiver Benutzung nicht vor.

Trotz des fortlaufenden Einsatzes von Fragebögen zur summativen Evaluation intuitiver Benutzung und der stetigen Bemühungen von einer Vielzahl von Forschern, den Administrationsaufwand und die Komplexität dieser Fragebögen zu reduzieren, bleiben eine Reihe von Limitationen der subjektiven Erfassung von intuitiver Benutzung bestehend (Blackler et al., 2018). So bleiben überwiegend methodische Limitationen, die verhindern, dass subjektive Methoden als Benchmark für die summative Evaluation intuitiver Benutzung im Rahmen des Projekts 3D-GUIde fungieren können. Eine Limitation bildet hierbei beispielsweise die Tatsache, dass Fragebögen in den meisten Fällen vom Nutzer erst nach erledigter Aufgabe (z.B. SMEQ bzw. SEA-Skala) oder nach der gesamten Systemnutzung (z.B. QUESI) retrospektiv ausgefüllt werden. Auf diese Weise können diese nicht die intuitive Benutzung im zeitlichen Verlauf der Systeminteraktion abbilden (d.h. keine kontinuierliche Abbildung intuitiver Benutzung). Lässt man beispielsweise die Nutzer während der Systeminteraktion das Ausmaß ihrer intuitiven Benutzung parallel mithilfe eines Fragebogens bewerten, kann dies ihre Aufgabenbearbeitung beeinflussen, da die zusätzliche Aufgabe ihre mentale Beanspruchung erhöhen kann (Reinhardt et al., 2018). Eine derartige Intrusion kann die wissenschaftliche Güte eines jeden Fragebogens gefährden (Eggemeier, Wilson, Kramer, & Damos, 1991).

Zusätzlich ist bei der Verwendung von Fragebögen zur retrospektiven Erfassung intuitiver Benutzung darauf hinzuweisen, dass Eindrücke zu Beginn oder zum Ende der Interaktion überproportional gewichtet werden (Blackler, 2018; Tretter et al., 2018). Laut Blackler (2018) ist dieser Primacy-Recency-Effekt ein typisches Beispiel für eine kognitive Verzerrung in der menschlichen Wahrnehmung (siehe Mayo & Crockett, 1964). Tretter et al. (2018) sehen die bei intuitiver Benutzung beteiligten unbewussten Prozesse und die retrospektive Natur von Maßen subjektiver Selbstbeurteilung für dieses Phänomen und die damit verbundene Ungenauigkeit verantwortlich. Derartige Wahrnehmungsfehler, von denen es viele gibt, gefährden die wissenschaftliche Güte des Messinstruments (Döring & Bortz, 2016). Eine Reihe weiterer empirischer Arbeiten in (z.B. Palmer, 2018; M. L. Still & Still, 2018) und außerhalb der Forschung zu intuitiver Benutzung (z.B. Cockburn,

Quinn, & Gutwin, 2015; Dell, Vaidyanathan, Medhi, Cutrell, & Thies, 2012) konnten bereits aufzeigen, dass eine Reihe von weiteren Faktoren wie beispielsweise Einstellungen und kulturelle Hintergründe subjektive Maße verzerren können. Aufgrund der genannten Einschränkungen reichen subjektive Maße für die summative Evaluation intuitiver Benutzung aus Perspektive des Projekts 3D-GUIde nicht aus, da lediglich die gesamte Systemnutzung oder einzelne Aufgaben bewertet werden können. Die für die Entwicklung von 3D-CUI-Interaktionspatterns kritischen einzelnen Interaktionslösungen können mit diesen zu „groben“ Maßen nicht evaluiert werden, da hierfür intuitive Benutzung kontinuierlich während der eigentlichen Systemnutzung bewertet werden muss. Blackler et al. (2018) raten aufgrund der Einschränkungen von subjektiven Methoden, zusätzlich auch objektive Methoden für die summative Evaluation intuitiver Benutzung einzusetzen, da nur diese als Benchmarks für die summative Evaluation intuitiver Benutzung in Frage kommen können.

Obwohl subjektive Methoden nicht alleine für die summative Evaluation von 3D-CUI-Interaktionslösungen im Rahmen des Projekts 3D-GUIde genutzt werden können und damit keine subjektive Methode als summativer Benchmark für die summative Evaluation intuitiver Benutzung in Frage kommt, können subjektive Methoden als wertvolle Quasi-Außenkriterien für die Sicherstellung der wissenschaftlichen Güte von neu entwickelten Methoden im Rahmen einer Meta-Evaluation fungieren. Die Maße SMEQ und NASA-TLX können hier bei der Approximation eines echten Außenkriteriums als Quasi-Außenkriterien genutzt werden, obwohl die wissenschaftliche Güte des NASA-TLX noch nicht formal im Bereich intuitiver Benutzung empirisch nachgewiesen wurde. Jedoch stellt dieser den Goldstandard zur Erfassung von mentaler Beanspruchung in der HCI dar. Darüber hinaus eignet sich der QUESI neben dem SMEQ und dem NASA-TLX als weiteres Quasi-Außenkriterium, da er gegenüber dem INTUI das Gefühl von Flüssigkeit anhand einer höheren Anzahl an Gefühlen approximiert. Außerdem ist er basierend auf aktuellen Reviews auch weiterverbreitet (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018). Zusätzlich ist seine Auswertung und Interpretation einfacher, da er hierfür einen Gesamtwert (d.h. gemittelter Wert über alle Skalen) anstelle einer komplexen Pattern-Interpretation vorsieht. Zusammenfassend kann mithilfe der Quasi-Außenkriterien SMEQ, NASA-TLX und QUESI ein echtes Außenkriterium auf Basis subjektiver Methoden approximiert und dieses für eine Meta-Evaluation einer neuen summativen Evaluationsmethode für intuitive Benutzung genutzt werden.

3.6.2 Objektive Methoden

Laut eines aktuellen Reviews von Blackler et al. (2018) wird zur objektiven Bewertung intuitiver Benutzung überwiegend auf objektive Daten zurückgegriffen, die in realen Nutzertests mit dem tatsächlichen System oder Prototypen gesammelt werden. Solche Beobachtungs- oder Verhaltensmaße, wie sie von Blackler et al. (2018) genannt werden, liefern wertvolle Informationen während der eigentlichen Systemnutzung, die aufgrund der retrospektiven Natur von subjektiven Maßen alleine mit diesen nicht zugänglich sind. Basierend auf dem aktuellsten Review von Blackler et al. (2018) können bei den Verhaltensmaßen zwei grundlegende Ansätze identifiziert werden, welche sich auch in älteren Reviews des Feldes wiederfinden lassen (z.B. Blackler & Popovic, 2015).

Der erste Ansatz versucht, die Richtigkeit und Differenziertheit des von Nutzern mitgebrachten Vorwissens für die durchzuführende Aufgabe im Verhältnis zum vom System erwarteten Vorwissen mithilfe von Image Schemas zu quantifizieren. Auf diese Weise lässt sich die Wahrscheinlichkeit abschätzen, wie stark unbewusst die spätere Interaktion mit dem System ausfällt. Aufgrund der Tatsache, dass die überwiegend unbewusste kognitive Informationsverarbeitung ein Korrelat des zentralen objektiven Merkmals intuitiver Benutzung, der mentalen Beanspruchung darstellt, wird intuitive Benutzung bei diesem Ansatz indirekt anhand der Korrelate dieses zentralen Merkmals objektiv beurteilt. Beim anderen Ansatz werden Nutzer bei der Systeminteraktion per Video aufgezeichnet (d.h. Videoaufnahme eines klassischen Nutzertests) und deren Interaktion retrospektiv von Experten anhand mehrerer Korrelate (z.B. zeitlich effiziente Informationsverarbeitung, überwiegend unbewusste kognitive Informationsverarbeitung) des objektiven zentralen Merkmals intuitiver Benutzung, der mentalen Beanspruchung, beurteilt.

Wie zuvor bei den subjektiven Methoden beschrieben, können Methoden, die ursprünglich für die Messung mentaler Beanspruchung entwickelt wurden, auch für die Evaluation intuitiver Benutzung nützlich sein. Eine Reihe von empirischen Befunden legt nahe, dass verschiedene physiologische Reaktionen (z.B. Hautleitfähigkeit, Herzratenvariabilität, Pupillenerweiterung) stark mit dem Ausmaß mentaler Beanspruchung zusammenhängen (Cain, 2007; Charles & Nixon, 2019; F. Chen et al., 2016; Cowley et al., 2016; Meshkati & Hancock, 2011; Young et al., 2015). Anhand dieser Reaktionen lassen sich Änderungen in der mentalen Beanspruchung objektiv direkt erkennen und auf dieser Basis die intuitive Benutzung des Systems beurteilen (Reinhardt & Hurtienne, 2017).

Die Kategorie der Hauptaufgabenleistungsmaße fußt auf der Annahme, dass sich die mentale Beanspruchung auf die Performance (z.B. Effektivität, Effizienz) des Nutzers objektiv auswirkt (Cain, 2007; Longo, 2014, 2017; Rubio et al., 2004; Tsang, 2006; Young et al., 2015). Aufgrund der Tatsache, dass sich intuitive Benutzung wiederum anhand der mentalen Beanspruchung beurteilen lässt, die Effektivität bei der Handlungsregulation stets ein Leistungskriterium für eine intuitive Handlung bildet und deswegen schon aus pragmatischen Gründen als weiteres zentrales Merkmal intuitiver Benutzung in der im Rahmen dieser Arbeit aufgestellten Arbeitsdefinition enthalten ist (siehe Teilabschnitt 2.2.2), werden in der Forschung zu intuitiver Benutzung Hauptaufgabenleistungsmaße wie die Korrektheit bei der Aufgabenbearbeitung und die Anzahl kritischer Fehler berücksichtigt (siehe Blackler & Popovic, 2015; Blackler et al., 2018; Hurtienne, 2011).

Die letzte Kategorie der Zweitaufgabenleistungsmaße nutzt den Umstand, dass mentale Ressourcen, die von der eigentlichen Systemnutzung (d.h. Hauptaufgabe) nicht gebunden werden, theoretisch noch verfügbar sind und deswegen für andere Aktivitäten eingesetzt werden können (Cain, 2007; F. Chen et al., 2016; Longo, 2014, 2017; Young et al., 2015). Das Zweitaufgabenleistungsmaß berechnet sich in diesem Zusammenhang als das Ausmaß der mentalen Ressourcen, die in der Zweitaufgabe gebunden werden. Aus der Leistung bei der Zweitaufgabe (z.B. Effektivität, Effizienz) kann deswegen die mentale Beanspruchung objektiv gemessen werden, die durch die Hauptaufgabe erzeugt wird und dadurch Aussagen bezüglich des Ausmaßes der intuitiven Systemnutzung getroffen werden (Reinhardt & Hurtienne, 2017).

Die oben genannten objektiven Methoden werden nun detailliert vorgestellt und bezüglich ihrer wissenschaftlichen Güte bewertet und Limitationen diskutiert. Darüber hinaus soll

aufgezeigt werden, welche vorgestellten objektiven Methoden sich als Quasi-Außenkriterien für eine Meta-Evaluation im Bereich intuitiver Benutzung eignen. Aufgrund der Tatsache, dass der im Zuge dessen identifizierte summative Benchmark jedoch eine geringe zeitliche Anwendungseffizienz aufweist und im Projekt 3D-GUIde eine zeitliche anwendungseffiziente summative Evaluationsmethode benötigt wird, kann die Entwicklung einer neuen summativen Evaluationsmethode namens IntuiBeat-S gerechtfertigt werden. Diese Methode soll stellvertretend für den aktuellen summativen Benchmark im Projekt 3D-GUIde zur zeitlich effizienten summativen Evaluation eingesetzt werden.

3.6.2.1 Verhaltensmaße

Q: Evaluation intuitiver Benutzung basierend auf Image Schemas

Wie bereits in Kapitel 2 angesprochen, ist das mentale Modell des Nutzers unter anderem durch Schemas strukturiert. Bei einer Art dieser Schemas handelt es sich um Image Schemas, welche abstrakte Repräsentationen von wiederkehrenden sensomotorischen Erfahrungen mit der Welt darstellen (Hurtienne, 2011, 2017; Hurtienne & Blessing, 2007; Löffler et al., 2013). Da Image Schemas das Vorwissen des Nutzers strukturieren und somit das Ausmaß potentieller intuitiver (d.h. überwiegend unbewusst) kognitiver Informationsverarbeitung beeinflussen, kann alleine die Berücksichtigung von Image Schemas bei der Gestaltung von Systemen zur Verbesserung der intuitiven Benutzung des Systems beitragen (Hurtienne, 2017; Löffler et al., 2013). Denn auf diese Weise kann eine geringe mentale Beanspruchung bei der Nutzung erreicht werden. Dazu müssen zunächst geeignete metaphorische Erweiterungen für die Image Schemas gefunden und so image-schematische Metaphern (d.h. Primärmetaphern) identifiziert werden. Betrachtet man beispielsweise die Ausdrücke „Die Aktienkurse gehen durch die Decke“ und „Der Mitarbeiter ist hochqualifiziert“, wird klar dass in diesen image-schematischen Metaphern die Quelldomäne aus einer vertikalen Dimension besteht, die vom zugrunde liegenden Image Schema UP - DOWN vorgegeben wird, das eine abstrakte Zieldomäne (d.h. Quantität) abbildet. Diesen Beispielen liegt die metaphorische Erweiterung MORE IS UP - LESS IS DOWN zugrunde, welche auf dem Image Schema UP - DOWN beruht (siehe Hurtienne, 2017).

Um festzustellen, welche dieser image-schematischen Metaphern die Nutzung eines bestimmten Systems am besten charakterisieren, wird beispielsweise der Nutzer im Rahmen einer Contextual Inquiry (siehe Holtzblatt & Beyer, 2016) bei seiner direkten Systemnutzung interviewt und basierend auf diesem Gespräch ein verbales Protokoll durch Transkription angefertigt. Basierend auf den dort enthaltenen Metaphern kann das System *metaphern-konform* gestaltet werden (Löffler et al., 2013). Muss beispielsweise entschieden werden, wie die Lautstärkeregelung durch einen Schieberegler realisiert werden kann, hilft dem User Interface Designer die image-schematische Metapher MORE IS UP - LESS IS DOWN bei der Gestaltung des Widgets (d.h. Schieberegler muss nach oben geschoben werden, um die Lautstärke zu erhöhen und nach unten für Senken der Lautstärke). Eine Reihe von wissenschaftlichen Arbeiten konnte bereits aufzeigen, dass Systeme, die metaphernkonform gestaltet werden zu einer höheren intuitiven Benutzung des Nutzers führen, als wenn die Gestaltung nicht konsistent mit den identifizierten image-schematischen Metaphern verlief. Eine detaillierte Übersicht dieser Arbeiten liefert Hurtienne (2017).

Basierend auf diesem Ansatz entwickelte Asikhia (2015) eine summative Evaluationsmethode, die das Ausmaß der intuitiven Benutzung eines Systems objektiv durch Experten anhand der Metaphernkonformität bestimmt. Hierzu muss zunächst das Ziel der Systemnutzung (z.B. Erfüllung einer bestimmten Aufgabe) identifiziert und durch Experten in eine Abfolge von Teilzielen aufgegliedert werden. Auf der untersten Ebene müssen dann für alle Teilziele die verschiedenen tatsächlichen Lösungswege festgehalten werden (d.h. welche Bedienelemente müssen für die Erreichung des Teilziels verwendet werden).

Als nächsten Schritt müssen die image-schematischen Metaphern identifiziert werden, die das Vorwissen des Nutzers und damit sein mentales Modell bei der Handlungsregulation beschreiben. Die image-schematischen Metaphern stammen dabei aus verschiedenen Quellen, wie unter anderem der Transkription des Contextual Interviews oder eines während der Systemnutzung parallel ausgeführten Think-Aloud-Protokolls. Als nächster Schritt werden die image-schematischen Metaphern extrahiert, die das zu evaluierende System beschreiben und somit das mentale Modell des User Interface Designers widerspiegeln. Die image-schematischen Metaphern werden dementsprechend aus den Funktions- und Aufgabenbeschreibungen aus dem Nutzerhandbuch des Systems entnommen. Zur Beurteilung des Ausmaßes intuitiver Benutzung muss abschließend der Grad der Überlappung von verwendeten image-schematischen Metaphern von Nutzer und User Interface Designer ermittelt werden. Asikhia (2015) konnte in einer ersten Pilotstudie bei der Evaluation von Weckern (d.h. drei Wecker mit unterschiedlichen Funktionen) zeigen, dass je höher die Überlappung von image-schematischen Metaphern zwischen User Interface Designer und Nutzer ist, umso höher auch das Ausmaß an intuitiver Benutzung des Nutzers ist. Der Grad der Überlappung, welche auch von Asikhia (2015) als Q-Wert bezeichnet wird, kann somit zur summativen Evaluation intuitiver Benutzung verwendet werden. Für die genaue Berechnung des Q-Wertes sei auf Asikhia (2015) verwiesen, ebenso für eine detailliertere Beschreibung der Quellen der image-schematischen Metaphern, eine detailliertere Diskussion des Vorgehens bezüglich der Metaphernextraktion und deren Analyse, da Image Schemas nicht im Zentrum dieser Arbeit stehen.

Asikhia (2015) demonstrierte die Objektivität seiner Methode, indem er deren Auswertungsobjektivität anhand der Beurteilerübereinstimmung bei der Extraktion der image-schematischen Metaphern nachwies. Das Cohens κ lag im Schnitt bei .59, was laut J. Cohen (1988) als moderat gilt. Die Durchführungs- und Interpretationsobjektivität wurde von Asikhia (2015) nicht empirisch gezeigt, was aber, wie bereits mehrfach erwähnt, kein Problem darstellt, sofern Reliabilität und Validität nachgewiesen sind (Lienert & Raatz, 1998). Reliabilität wurde von Asikhia (2015) nicht empirisch untersucht. Die Validität seiner Methode als objektives Maß für intuitive Benutzung zeigte Asikhia (2015) in Form der inneren kriteriumsbezogenen Validität, indem er bei der Evaluation auch zusätzlich die mentale Beanspruchung der Nutzer mithilfe einer eigenen nicht standardisierten Skala erfasste und anschließend mit dem dazugehörigen Q-Wert verglich (Wecker 1: r = -.70; Wecker 2: r = -.77; Wecker 3: r = -.62). Laut J. Cohen (1988) können die Korrelationen als hoch (d.h. $|r|$ um .50) interpretiert werden. Darüber hinaus erfasste Asikhia (2015) noch die Effektivität anhand der Fehleranzahl und konnte auch dabei unter Berücksichtigung von J. Cohen (1988) vergleichsweise hohe Korrelationen (d.h. $|r|$ um .50) mit dem Q-Wert der einzelnen Wecker feststellen (Wecker 1: r = -.57; Wecker 2: r = -.64; Wecker 3: r = -.44). Schließlich korrelierte Asikhia (2015) die Q-Werte der Wecker mit der Zeit der Interaktion und stellte auch dort in Übereinstimmung mit J. Cohen (1988) hohe Zu-

sammenhänge (d.h. $|r|$ um .50) fest (Wecker 1: r = -.70; Wecker 2: r = -.77; Wecker 3: r = -.62).

Da Asikhia (2015) seine Q-Methode nur an konstruktnahen und nicht auch an konstruktfernen Kriterien evaluiert hat, ist nur ein erster empirischer Nachweis bezüglich einer Kriteriumsvalidierung verfügbar. Man kann daher nicht von einer Konstruktvalidierung sprechen. Darüber hinaus muss an dieser Stelle noch erwähnt werden, dass die Methode bisher nicht außerhalb der Dissertation von Asikhia (2015) zum Einsatz kam. Obwohl es mit der Arbeit von Asikhia (2015) einen ersten Beleg für die wissenschaftliche Güte des Instruments gibt, ist seine praktische Güte unter Berücksichtigung der Anforderung des 3D-GUIde Projekts nach einer zeitlich effizienten Methode fragwürdig, da die von Asikhia (2015) vorgeschlagene Auswertung von image-schematischen Metaphern zeitlich gesehen bereits bei einfachen System wie Weckern sehr aufwendig ist. Des Weiteren konnten weder Asikhia (2015) in seiner Arbeit noch aktuelle Reviews des Feldes (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) empirische Nachweise bezüglich der zeitlichen Anwendungseffizienz der Methode im Vergleich zu anderen objektiven Methoden liefern, weswegen der Methode keine hohe zeitliche Anwendungseffizienz attestiert werden kann.

CHAI: Coding Heuristics for Assessing Intuitive Interaction

Neben der Beurteilung der Richtigkeit und Differenziertheit des mentalen Modells durch Analyse der bei einem System vorliegenden Metaphernkonformität, kann intuitive Benutzung auch anhand mehrerer Korrelate des objektiven zentralen Merkmals intuitiver Benutzung (siehe Teilabschnitt 2.2.2), der mentalen Beanspruchung, von Experten objektiv beurteilt werden. Ein erstes Bewertungsschema für die Beurteilung verschiedener Korrelate (unter anderem zeitlich effiziente Informationsverarbeitung, überwiegend unbewusste kognitive Informationsverarbeitung) und der effektiven Benutzung stellte die QUT-Forschergruppe bereit (z.B. Blackler, 2006; Blackler et al., 2018; Blackler et al., 2011; Blackler et al., 2010; Lawry, Popovic, & Blackler, 2011). Die Anwendung dieses Schemas erfolgte meines Wissens auch überwiegend durch die QUT-Forschergruppe (siehe Blackler & Popovic, 2015; Blackler et al., 2018).

Um das QUT-Bewertungsschema anwenden zu können, müssen Nutzer zuvor bei der Systemnutzung per Video aufgezeichnet werden und ihre Systemnutzung währenddessen mithilfe eines parallelen Think-Aloud-Protokolls beschreiben (Blackler et al., 2018). Experten analysieren im Anschluss jede Interaktion innerhalb dieser Videos mithilfe eines Tools zur Verhaltensbeobachtung wie Noldus Observer XT (Zimmerman, Bolhuis, Willemsen, Meyer, & Noldus, 2009) oder BORIS (Friard & Gamba, 2016). Es eignen sich jedoch auch einfache Videoplayer, sofern man das Video auch Bild für Bild durchgehen kann (Reinhardt et al., 2018).

Mithilfe des QUT-Bewertungsschemas (siehe Abbildung 3.3) wird bei der Beurteilung einer jeden Interaktion zum einen die objektive mentale Beanspruchung (bei Blackler (2006) als *Art der Benutzung* bezeichnet) anhand von fünf auf Korrelaten intuitiver Benutzung basierenden Kriterien (d.h. zeitliche Effizienz bei der kognitiven Informationsverarbeitung: (1) Handlungsinitiierung innerhalb von 5 Sekunden bzw. Schnelligkeit, (2) zielgerichtete

Handlungsausführung bzw. Entscheidungssicherheit; überwiegend unbewusste kognitive Informationsverarbeitung auf Basis von handlungsrelevantem Vorwissen: (3) keine Verbalisierung der kognitiven Informationsverarbeitung während der Interaktion bzw. bewusste Verarbeitung, (4) klare Erwartungen aufgrund von vorhandenem handlungsrelevanten Vorwissen bzw. Erwartung, (5) klarer Bezug zu handlungsrelevantem Vorwissen bzw. Verbindung zu vergangenen Erfahrungen) beurteilt. Zum anderen wird die *Effektivität der Benutzung* anhand von fünf Kriterien (d.h. (1) „korrekt“, (2) „korrekt, aber nicht der Aufgabe angemessen“, (3) „inkorrekt“, (4) „versucht“, (5) „Hilfe erhalten“) zur Bestimmung des Ausmaßes an intuitiver Benutzung von Experten berücksichtigt. Um laut QUT-Forschergruppe als intuitiv benutzbar zu gelten, muss eine Interaktion effektiv, also „korrekt“ sein und mindestens zwei der fünf Kriterien zur Beurteilung mentaler Beanspruchung (d.h. Art der Benutzung) erfüllen.

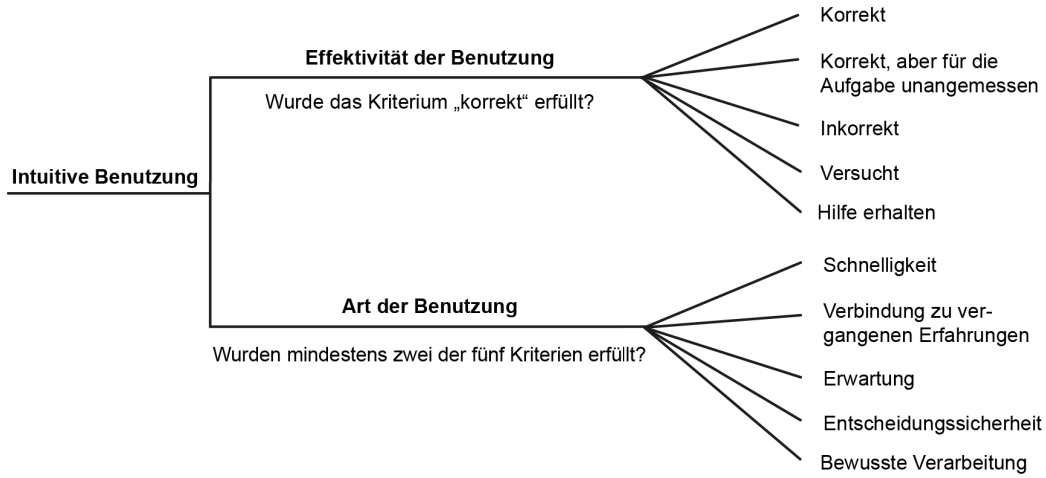


Abbildung 3.3. QUT-Bewertungsschema von Blackler (2006) in Anlehnung an Reinhardt, Kuge und Hurtienne (2018).

Die Reliabilität des QUT-Bewertungsschemas wurde in den im Review genannten Arbeiten nicht thematisiert (siehe Blackler et al., 2018). Auch frühere Reviews (siehe Blackler & Hurtienne, 2007; Blackler & Popovic, 2015), thematisierten diesen Aspekt nicht. Ebenso wenig wurde die Objektivität empirisch nachgewiesen und von Blackler (2006) nur theoretisch diskutiert.

Bezüglich der Validität kann der Methode von Blackler (2006) als objektives Maß für intuitive Benutzung zwar unter der Berücksichtigung der im zweiten Kapitel dieser Arbeit erarbeiteten Arbeitsdefinition (siehe Teilabschnitt 2.2.2), sowie der selbst von der QUT-Forschergruppe verwendeten Literatur (siehe Teilabschnitt 2.1.2) Inhaltsvalidität zugesprochen werden, Kriteriums- und Konstruktvalidität des Schemas sind aber als fragwürdig anzusehen. Blackler (2006) und Blackler et al. (2010) nahmen zur Sicherstellung der inneren kriteriumsbezogenen Validität beispielsweise den bereits vorgestellten TFQ als Quasi-Außenkriterium her, der selbst nicht von hoher wissenschaftlicher Güte ist (siehe Tabelle 3.2), was zu einem Zirkelschluss führen könnte. Auch bezüglich der praktischen Güte in Form von zeitlicher Anwendungseffizienz weist das QUT-Bewertungsschema große

Defizite auf. So fällt der zeitliche Kodierungsaufwand bereits bei Digitalkameras (siehe Blackler et al., 2010), die im Vergleich zu komplexen Systemen wie CUIs relativ wenige Interaktionen aufweisen, bereits sehr hoch aus. Die zeitliche Anwendungseffizienz der Methode als wichtiger Aspekt praktischer Güte lässt sich deswegen allgemein als gering einstufen. Aktuelle Reviews des Feldes (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) konnten dabei außerdem keine empirischen Nachweise bezüglich der zeitlichen Anwendungseffizienz der Methode im Vergleich zu anderen objektiven Methoden liefern, weswegen der Methode auch aus dieser Perspektive keine hohe zeitliche Anwendungseffizienz attestiert werden kann.

Das QUT-Bewertungsschema (siehe Abbildung 3.3) wurde insbesondere aufgrund seiner geringen zeitlichen Anwendungseffizienz und weiterer methodischer Mängel von einigen Wissenschaftlern kritisiert. Neben der reinen Anzahl an zu prüfenden Kriterien und der damit geringen Effizienz kritisierten Horn (2008) sowie Reinhardt et al. (2018) hauptsächlich die Auswahl der genutzten Kriterien an sich. So erfordern drei der fünf Kriterien zur Beurteilung der mentalen Beanspruchung ein paralleles Think-Aloud-Protokoll. Jedoch zeigten Hertzum et al. (2009), dass ein zur Handlung parallel durchgeführtes, verbales Think-Aloud-Protokoll zusätzliche mentale Beanspruchung beim Nutzer erzeugt, was dementsprechend die Beurteilung mentaler Beanspruchung und damit auch das gemessene Ausmaß intuitiver Benutzung stark verzerren kann (Reinhardt et al., 2018).

Aufgrund der genannten Limitationen bezüglich der wissenschaftlichen und zeitlichen Anwendungseffizienz des QUT-Bewertungsschemas kam es zur Entwicklung des Bewertungsschemas *CHAI* (Coding Heuristics for Assessing Intuitive Interaction) (siehe Abbildung 3.4), welches eine Überarbeitung des ursprünglichen QUT-Bewertungsschemas der QUT-Forschergruppe zur Kompensation seiner Defizite darstellt (Horn, 2008; Reinhardt et al., 2018). Das CHAI-Bewertungsschema kommt im Vergleich zum QUT-Bewertungsschema mit insgesamt weniger Kriterien aus (siehe Abbildung 3.4) und ist bei seiner Anwendung nicht auf ein möglicherweise kompromittierendes Think-Aloud-Protokoll angewiesen. Zur Senkung des Kodierungsaufwands der Experten wurden beim CHAI-Bewertungsschema außerdem die Anzahl der Kriterien zur Beurteilung der Effektivität von fünf auf vier und zur Beurteilung der mentalen Beanspruchung während der Interaktion von fünf auf zwei reduziert. Bei der Formulierung dieser Kriterien konzentrierten sich Reinhardt et al. (2018) wie beim QUT-Bewertungsschema auf die Korrelate *zeitlich effiziente kognitive Informationsverarbeitung* und *überwiegend unbewusste Anwendung von Vorwissen während der Benutzung*, stellten aber pro Korrelat nur ein Kriterium bereit. Wie im Teilabschnitt 2.2.2 erwähnt, resultiert eine objektiv beurteilbare geringe mentale Beanspruchung immer aus einer hohen objektiven Flüssigkeit bei der kognitiven Informationsverarbeitung, weswegen daraus auch eine elegante, schnelle, akkurate und sparsame Bewegung des Nutzers resultiert (Graf et al., 2018; R. Reber et al., 2004; Slepian & Ambady, 2012; Topolinski & Strack, 2009).

Die Wahrscheinlichkeit einer intuitiven Interaktion ist hoch, wenn die Ausführungszeit einer Interaktion unter drei Sekunden bleibt (Horn, 2008; Reinhardt et al., 2018). Diese Zahl wurde basierend auf dem Keystroke-Level-Model für die Interaktion mit Desktopsystemen berechnet. Das Keystroke-Level-Model ist eine einfache Technik zur kognitiven Modellierung, die dazu verwendet werden kann, abzuschätzen wie lange Experten für die Durchführung einer Routineaufgabe benötigen (Card, Moran, & Newell, 1980). Die

Abschätzung erfolgt dabei auf Basis von generalisierten atomaren, kognitiven und motorischen Operationen, für die empirisch bestimmte Standardzeiten vorliegen (Horn, 2008; Reinhardt et al., 2018). Dieses erste Kriterium zur objektiven Beurteilung der durch eine Interaktion verursachten mentalen Beanspruchung wird beim CHAI-Bewertungsschema als *Schnelligkeit* bezeichnet.

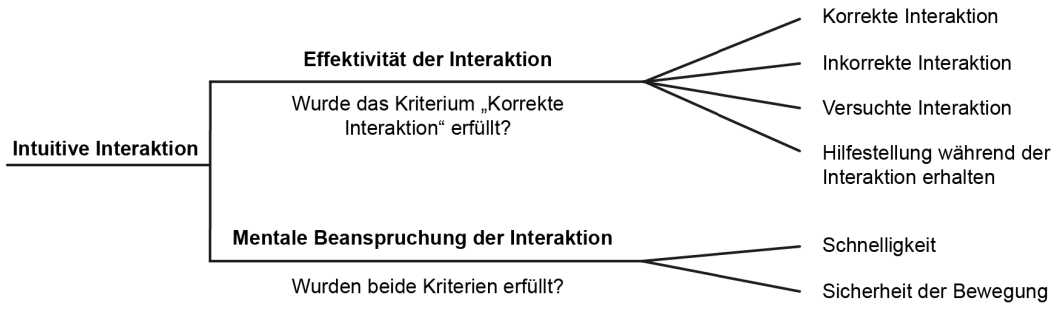


Abbildung 3.4. CHAI-Bewertungsschema in Anlehnung an Reinhardt, Kuge und Hurtienne (2018).

Das zweite Kriterium des CHAI-Bewertungsschemas zur objektiven Bewertung der mentalen Beanspruchung wird als *Sicherheit der Bewegung* bezeichnet und beurteilt wie akkurat die Interaktionsbewegung an sich war, da eine hohe objektive Flüssigkeit bei der kognitiven Informationsverarbeitung auch zu besonders akkuraten Bewegungen des Nutzers führt (Graf et al., 2018; R. Reber et al., 2004; Slepian & Ambady, 2012; Topolinski & Strack, 2009). Führt eine Bewegung ohne Unterbrechungen, längerem Suchen oder allgemeinen Kreisbewegungen zum Interaktionsziel, desto höher ist die Wahrscheinlichkeit und das Ausmaß intuitiver Benutzung (Horn, 2008; Reinhardt et al., 2018).

Neben den Kriterien zur leitfadenorientierten Einschätzung der mentalen Beanspruchung wurden, wie bereits erwähnt, auch die im QUT-Bewertungsschema enthaltenen fünf Kriterien zur *Beurteilung der Effektivität* (d.h. „korrekt“, „korrekt, aber nicht der Aufgabe angemessen“, „inkorrekt“, „versucht“, „Hilfe erhalten“) auf die folgenden vier Abstufungen reduziert (Horn, 2008; Reinhardt et al., 2018):

Korrekte Interaktion Eine Interaktion ist als korrekt zu bewerten, wenn mit dem beteiligten Steuerelement (z.B. Button) erwartungsgemäß interagiert wurde und die Interaktion zielführend bezüglich der Aufgabe oder Teilaufgabe war.

Inkorrekte Interaktion Eine Interaktion ist als inkorrekt zu bewerten, wenn mit dem beteiligten Steuerelement (z.B. Button) nicht erwartungsgemäß interagiert wurde und/oder die Interaktion nicht zielführend bezüglich der Aufgabe oder Teilaufgabe war.

Versuchte Interaktion Eine Interaktion ist als versucht zu bewerten, wenn die Interaktion aufgrund eines Systemfehlers nicht registriert werden konnte.

Hilfestellung während der Interaktion erhalten Eine Interaktion unter Hilfestellung liegt dann vor, wenn der Nutzer entweder innerhalb des Systems (z.B. Hilfefunktion) oder außerhalb um Hilfe (z.B. Versuchsleiter) ersucht hat.

Gemäß des CHAI-Bewertungsschemas ist eine Interaktion als intuitiv zu bewerten, wenn diese von einem Experten als effektive Interaktion eingestuft wurde und gleichzeitig dessen geringe mentale Beanspruchung durch die Erfüllung der beiden damit verbundenen Kriterien bestätigt werden konnte. Horn (2008) führte eine erste Meta-Evaluation des CHAI-Bewertungsschemas bei Fahrkartenautomaten durch, wobei sie jeden Knopfdruck im System als eine abgeschlossene Interaktion wertete. Der Anteil der intuitiven Interaktionen (d.h. Intuitiv-Klicks) an der Gesamtzahl an Interaktionen (d.h. Gesamtklicks) kam als Operationalisierung des Ausmaßes intuitiver Benutzung bei der summativen Evaluation zum Einsatz.

Die Objektivität der CHAI-Methode wurde zur Sicherstellung der wissenschaftlichen Güte von Horn (2008) nicht explizit empirisch aufgegriffen, ebenso wenig wie die Reliabilität. Horn (2008) erwähnte in ihrer Arbeit jedoch kurz, dass sich bei der Kodierung der Intuitiv-Klicks nicht zufriedenstellende Beurteilerübereinstimmungen ergeben haben ohne dabei jedoch explizite Angaben zu deren genauen Höhe zu machen. Nichtsdestotrotz lieferte Horn (2008) Erkenntnisse bezüglich der Validität des CHAI-Beurteilungsschemas als objektives Maß intuitiver Benutzung in Form von Konstruktvalidität. Zur Sicherstellung der konvergenten Konstruktvalidität korrelierte Horn (2008) hierzu den Anteil von Intuitiv-Klicks an Gesamtklicks mit dem subjektiven Benchmark QUESI (d.h. Wahrgenommene mentale Beanspruchung: $r = .20$; Wahrgenommene Zielerreichung: $r = .29$; Wahrgenommene Fehlerrate: $r = .21$; Wahrgenommener Lernaufwand: $r = .25$; Wahrgenommene Vertrautheit: $r = .19$; QUESI-Gesamtwert: $r = .41$). Laut J. Cohen (1988) können diese Zusammenhänge als mittel (d.h. $|r|$ um .30) eingestuft werden. Als zusätzlichen Nachweis für die konvergente Validität der CHAI-Methode korrelierte sie den Anteil an Intuitiv-Klicks mit dem Anteil korrekt gelöster Aufgaben (d.h. Effektivität) ($r = .60$) und der Zeit für die Interaktion (d.h. zeitliche Effizienz bei der kognitiven Informationsverarbeitung) ($r = .62$). Beide Zusammenhänge können laut J. Cohen (1988) als hoch (d.h. $|r|$ um .50) interpretiert werden. Die divergente Validität der Methode wies Horn (2008) dadurch nach, dass sie den Anteil an Intuitiv-Klicks mit der Anzahl der Gesamtklicks (d.h. physische Effizienz bei der Handlungsausführung) korrelierte ($r = 22$, *n. s.*). Aufgrund der Tatsache, dass sowohl die konvergente als auch die divergente Validität nachgewiesen wurden und die Absolutwerte der konvergenten Korrelationskoeffizienten höher als die Absolutwerte der divergenten Korrelationskoeffizienten lagen (siehe Moosbrugger & Kelava, 2007), kann der CHAI-Methode Konstruktvalidität attribuiert werden.

Reinhardt et al. (2018) konnten im Gegensatz zu Horn (2008) die Auswertungsobjektivität der Methode durch Berechnung einer Beurteilerübereinstimmung bei der Evaluation des CUIs *SketchUp* von Trimble Navigation Ltd. (2016) empirisch nachweisen, da die Auswertungsobjektivität der Methode aufgrund der Anpassungen des QUT-Bewertungsschemas von den Wissenschaftlern als besonders kritisch eingestuft wurde. Dabei wurde jeder Mausklick als eine abgeschlossene Interaktion angesehen. Der Anteil der intuitiven Interaktionen (d.h. Intuitiv-Klicks) an der Gesamtzahl an Interaktionen (d.h. Gesamtklicks) kam als Operationalisierung des Ausmaßes intuitiver Benutzung zur summativen Evaluation zum Einsatz. Die Beurteilerübereinstimmung zeigte ein Cohens κ von .61, was laut Field (2017) als eine hohe Übereinstimmung einzustufen ist. Durchführungs- und Interpretationsobjektivität wurden von Reinhardt et al. (2018) nicht thematisiert. Auch empirische Ergebnisse bezüglich der Reliabilität fehlen bei Reinhardt et al. (2018). Da jedoch die

Prüfung der Validität in gewissem Maße von der Überprüfung der übrigen Gütekriterien entbindet (Lienert & Raatz, 1998), stellt dieser Umstand kein größeres Problem dar.

Die Konstruktvalidität der CHAI-Methode als objektives Maß für intuitive Benutzung konnte ebenfalls von Reinhardt et al. (2018) nachgewiesen werden. Zur Bestätigung der konvergenten Validität korrelierten sie den Anteil an Intuitiv-Klicks mit dem subjektiven Quasi-Außenkriterium SMEQ (in Form der deutschen SEA-Skala), welcher die mentale Beanspruchung erfasst, und stellten dabei laut J. Cohen (1988) eine hohe Korrelation (d.h. $|r|$ um .50) fest (r = -.61). Außerdem konnten sie im Gegensatz zu Horn (2008) auch einen hohen Zusammenhang (d.h. $|r|$ um .50) unter Berücksichtigung von J. Cohen (1988) zwischen dem Anteil an Intuitiv-Klicks und dem subjektiven Quasi-Außenkriterium QUESI feststellen (d.h. Wahrgenommene mentale Beanspruchung: r = .61; Wahrgenommene Zielerreichung: r = .61; Wahrgenommene Fehlerrate: r = .65; Wahrgenommener Lernaufwand: r = .65; Wahrgenommene Vertrautheit: r = .60; QUESI-Gesamtwert: r = .67). Zur Bestätigung der divergenten Validität korrelierten sie, wie Horn (2008) zuvor, den Anteil an Intuitiv-Klicks mit der Anzahl an Gesamtklicks (r = -.32, *n. s.*). Dementsprechend konnte von Reinhardt et al. (2018) die Konstruktvalidität unabhängig der von Horn (2008) getesteten Fahrkartenautomaten bei einem weitaus komplexeren User Interface (d.h. einem 3D-CUI) empirisch nachgewiesen werden. Bei der Studie von Reinhardt et al. (2018) soll jedoch angemerkt werden, dass die Meta-Evaluation auf Basis der im Rahmen dieser Arbeit erhobenen Daten des ersten Experiments erfolgte (siehe Abschnitt 5.1).

Sowohl Horn (2008) als auch Reinhardt et al. (2018) konnten jedoch keinen empirischen Nachweis bezüglich der zeitlichen Anwendungseffizienz der Methode im Vergleich zu anderen objektiven Methoden liefern, weswegen der Methode keine hohe zeitliche Anwendungseffizienz attestiert werden kann, sie aber aufgrund der Reduktion der Anzahl benötigter Beurteilungskriterien im Vergleich zum ursprünglichen QUT-Bewertungsschema eine höhere zeitliche Anwendungseffizienz verspricht.

3.6.2.2 Physiologische Maße

Maße der Herz- und Kreislaufaktivität

Die mentale Beanspruchung lässt sich physiologisch anhand der Herz- und Kreislaufaktivität erfassen. Beispielsweise können mithilfe eines Elektrokardiogramms (EKG) die elektrischen Aktivitäten aller Herzmuskelfasern aufgezeichnet werden, welche bei den Kontraktionen des Herzmuskels entstehen. Aus dem EKG kann die *Herzfrequenzratenvariabilität (HRV)* abgeleitet werden, was das populärste Maß der Herz- und Kreislaufaktivität zur Abbildung mentaler Beanspruchung darstellt und aus der Perspektive von einer Reihe von Wissenschaftlern eine hohe Augenscheinvalidität besitzt (z.B. Aasman, Mulder, & Mulder, 1987; Backs, 1995; Brookhuis & De Waard, 2010; Charles & Nixon, 2019; Cowley et al., 2016; Eggemeier et al., 1991; Galy et al., 2012; L. J. Mulder, De Waard, & Brookhuis, 2004; Paxion, Galy, & Berthelon, 2014; Tattersall & Hockey, 1995; Young et al., 2015).

Die HRV bildet die Variabilität des RR-Intervalls ab (Charles & Nixon, 2019; F. Chen et al., 2016; Cowley et al., 2016). Als RR-Intervall wird die zeitliche Differenz zwischen

dem Beginn zweier Kontraktionen der Herzkammern, welche im EKG als Peaks oder sogenannte R-Zacken erscheinen, bezeichnet (Cowley et al., 2016; Paxion et al., 2014). Als Indikatoren können die mittlere Herzfrequenz (d.h. Anzahl an Schlägen pro Minute) und eine differentielle Herzfrequenz verwendet werden, um den Unterschied zwischen einer Ruhe- und einer Aktivitätsperiode zu identifizieren (Charles & Nixon, 2019; Cowley et al., 2016). Die differentielle Herzfrequenz quantifiziert damit die Standardabweichung der RR-Intervalle. Außerdem wird das EKG durch eine Analyse der spektralen Leistungsdichte, was üblicherweise mithilfe einer Fourier-Transformation durchgeführt wird, in mehrere Frequenzbänder unterteilt, um die HRV anhand ihres Spektrums analysieren zu können (Charles & Nixon, 2019; Cowley et al., 2016). Hier zeigt sich insbesondere das Zentrum des mittleren Frequenzbandes (d.h. 0.10 Hz-Komponente) für die Evaluation mentaler Beanspruchung als besonders geeignet (Charles & Nixon, 2019; Cowley et al., 2016; De Waard, 1996; Paxion et al., 2014; Veltman & Gaillard, 1998). Laut De Waard (1996) und Veltman und Gaillard (1998) zeigt sich bei steigender mentaler Beanspruchung eine Erhöhung der Herzfrequenzratenvariabilität im Zeitbereich und eine Senkung der Leistung im Frequenzbereich.

Die wissenschaftliche Güte der HRV als objektives Maß für mentale Beanspruchung wurde in den von Cowley et al. (2016) und Charles und Nixon (2019) genannten Arbeiten (z.B. Delaney & Brodie, 2000; Ryu & Myung, 2005; Tattersall & Hockey, 1995) überwiegend durch die in Teilabschnitt 3.2.5 vorgestellten nicht formalen Gütekriterien (insbesondere anhand der Kriterien *Sensitivität* und *Diagnostizität*) sichergestellt, welche für einen vollständigen Nachweis der wissenschaftlichen Güte nicht ausreichend sind (Hancock & Matthews, 2019; Matthews et al., 2015). Da im Rahmen dieser Arbeit nur Evaluationsmethoden für die summative Evaluation intuitiver Benutzung berücksichtigt werden, deren wissenschaftliche Güte formal anhand der psychometrischen Gütekriterien *Objektivität*, *Validität* und *Reliabilität* empirisch nachgewiesen wurde, werden im weiteren Verlauf dieses Teilabschnitts nur konkrete Ergebnisse bezüglich dieser Gütekriterien berichtet. Für konkrete Ergebnisse bezüglich nicht formaler Gütekriterien wird auf Charles und Nixon (2019) verwiesen. Aufgrund des mangelnden formalen Gütenachweises attestieren eine Reihe von Experten der HRV höchstens Inhaltsvalidität und zweifeln die Erfüllung der formalen Hauptgütekriterien an (z.B. Apparies, Riniolo, & Porges, 1998; Caldwell, Wilson, & Cetinguc, 1994; Manzey, 1997; Nickel & Nachreiner, 2003; Paas & Van Merriënboer, 1994; Veltman & Gaillard, 1996; G. F. Wilson, 1992). Die HRV kommt daher auch nicht als eine objektiv summative Evaluationsmethode für intuitive Benutzung im Rahmen des Projekts 3D-GUIde in Frage oder eignet sich als Quasi-Außenkriterium für eine Meta-Evaluation einer summativen Evaluationsmethode für intuitive Benutzung.

Es wird bezüglich der wissenschaftlichen Güte der HRV von vielen Arbeiten schwerpunktmäßig kritisiert, dass lediglich ein marginaler Zusammenhang zwischen mentaler Beanspruchung und der HRV besteht, weswegen feine Abstufungen (d.h. es kann damit nur zwischen hoher und geringer mentaler Beanspruchung unterschieden werden) bezüglich mentaler Beanspruchung damit auch kaum erkannt werden können (Charles & Nixon, 2019). Dies konnte bereits von einer Reihe von empirischen Arbeiten gezeigt werden (z.B. De Rivecourt, Kuperus, Post, & Mulder, 2008; Dussault, Jouanin, Philippe, & Guezennec, 2005; Y.-H. Lee & Liu, 2003; Splawn & Miller, 2013). Außerdem scheint sich das mittlere Frequenzband nicht zur Analyse von Aufgaben zu eignen, welche längere Sequenzen von motorischen Aktionen (z.B. Mausklicks, Tastatureingaben oder Sprechen) umfassen (De

Waard, 1996). Ferner reagiert das mittlere Frequenzband auch auf Änderungen im emotionalen Stressempfinden (z.B. Nickel & Nachreiner, 2003), in der Tageszeit (z.B. Schellekens, Sijtsma, Vegter, & Meijman, 2000) und in der Atmung (z.B. Bernardi et al., 2000) sensibel, was die wissenschaftliche Güte des Maßes zusätzlich gefährden kann (Charles & Nixon, 2019; Paxion et al., 2014). Auf Basis dieser und anderer Befunde (z.B. Braby, Harris, & Muir, 1993; G. Mulder, Mulder, Meijman, Veldman, & Van Roon, 2000) argumentieren Charles und Nixon (2019) in ihrem Review-Artikel, dass die HRV am besten nur für die Unterscheidung zwischen geringer und hoher mentaler Beanspruchung (d.h. Unterscheidung zwischen aktiven Phasen und Ruhephasen) angewendet werden soll. Für feinere Abstufungen fehlt der HRV die nötige Sensitivität und Diagnostizität. Beim Einsatz der HRV ist außerdem die Einsatzdomäne und die Dauer der Messung von besonderer Wichtigkeit (Charles & Nixon, 2019).

Eng mit der HRV verbunden ist der *Blutdruck (BD)* (d.h. Maß des ausgeübten Drucks auf die Blutgefäße aufgrund der Blutzirkulation im Körper), der üblicherweise entweder als der ausgeübte Druck durch das Zusammenziehen des Herzmuskels (d.h. systolischer Blutdruck) oder als der ausgeübte Druck durch das Entspannen des Herzmuskels (d.h. diastolischer Blutdruck) angegeben wird (Charles & Nixon, 2019). Gemessen werden kann der BD beispielsweise mithilfe eines digitalen Blutdruckmessgeräts (z.B. gemessen am Finger nach der Methode von Penaz (1973)). Dieses objektive Maß kommt laut aktuellen Reviews von Cowley et al. (2016) und Charles und Nixon (2019) sehr selten für die summative Evaluation von mentaler Beanspruchung zum Einsatz, da es sehr stark auf energetische, emotionale (d.h. emotionaler Stress, physische Anstrengung, siehe Jahn, Oehme, Krems, & Gelau, 2005) und respiratorische Prozesse (z.B. Bernardi et al., 2000) reagiert. Außerdem wird der BD unter anderem noch von Prozessen der Wärmeregulierung (z.B. Nickel & Nachreiner, 2003), der Tageszeit, der Verdauung, des vorhandenen Schlafs und vom Sprechen beeinflusst (z.B. Adams & Leverland, 1985). Des Weiteren wird für eine langfristige Messung mentaler Beanspruchung anhand des Blutdrucks im Vergleich zum EKG eine teurere Ausstattung benötigt, was sich auf die Kosten der Methode und somit auf die Effizienz der Methode auswirkt. Außerdem kann die Messung des Blutdrucks an einer Extremität (z.B. Finger oder Oberarm) den Nutzer leicht in seiner Systemnutzung behindern und kann damit als intrusiv bezeichnet werden (Cowley et al., 2016).

Bei der wissenschaftlichen Güte des BD als objektives Maß für mentale Beanspruchung wurde von Cowley et al. (2016) und Charles und Nixon (2019) nicht zwischen systolischem und diastolischem Blutdruck unterschieden, sondern unabhängig davon dessen Sicherstellung durch die in Teilabschnitt 3.2.5 vorgestellten nicht formalen Gütekriterien (insbesondere Sensitivität und Diagnostizität) diskutiert, die wie bereits erwähnt, für den Nachweis der wissenschaftlichen Güte der Methode nicht ausreichend sind (Hancock & Matthews, 2019; Matthews et al., 2015). Dementsprechend können BD wie HRV nicht als objektiv summative Evaluationsmethoden für intuitive Benutzung im Rahmen des Projekts 3D-GUIde berücksichtigt werden. Sie qualifizieren sich daher beide auch nicht als Quasi-Außenkriterien für eine Meta-Evaluation einer summativen Evaluationsmethode für intuitive Benutzung. Einige Forscher attestierten dem BD zwar auf Basis dieser nicht formalen Kriterien ausreichende wissenschaftliche Güte (z.B. Causse, Sénard, Démonet, & Pastor, 2010; Finsen, Søgaard, Jensen, Borg, & Christensen, 2001; Hjortskov et al., 2004; Madden & Savard, 1995; Ring, Burns, & Carroll, 2002; Stuiver, Brookhuis, De Waard, & Mulder, 2014; Veltman & Gaillard, 1996), wohingegen andere die wissenschaftliche Gü-

te des BD als eine objektiv summative Evaluationsmethode für mentale Beanspruchung aufgrund des fehlenden formalen empirischen Nachweises bezüglich der psychometrischen Gütekriterien, sowie der genannten Limitationen wie die Anfälligkeit gegenüber der Atmung anzweifeln (z.B. Bernardi et al., 2000; Boutcher & Boutcher, 2006; Hwang et al., 2008).

Neben der HRV und dem BD, kann die mentale Beanspruchung auch noch anhand der *Vasokonstriktion* (d.h. Volumenschwankungen der Gefäße) beurteilt werden, die in den aktuellen beiden Reviews von Cowley et al. (2016) und Charles und Nixon (2019) nur am Rande thematisiert wird. Mithilfe der Photoplethysmographie (PPG), bei der die Haut des Nutzers einer Infrarotstrahlung mit definierter Wellenlänge (d.h. 800 - 1000 Nanometer) ausgesetzt wird, lässt sich eine erhöhte mentale Beanspruchung durch eine Verengung der Blutgefäße feststellen, da das Hämoglobin die Strahlung wesentlich stärker absorbiert als das umgebende Gewebe. Nimmt bei erhöhter mentaler Beanspruchung die Blutmenge im subkutanen Venenplexus ab, so wird die Absorption geringer, die Reflexion höher. Der reflektierte Anteil und damit die Füllungsschwankungen im subkutanen Venenplexus werden durch die PPG gemessen und quantifizieren damit die mentale Beanspruchung des Nutzers. Für die genaue Funktionsweise diese Verfahrens sei an dieser Stelle jedoch auf Allen (2007) verwiesen, da sie wesentlich komplexer ist, als es hier dargestellt werden konnte. Laut der beiden Reviews existieren zwar einige empirische Hinweise (z.B. Bousefsaf, Maaoui, & Pruski, 2014; Iani, Gopher, Grunwald, & Lavie, 2007; Iani, Gopher, & Lavie, 2004; S. Miyake et al., 2009), die auf die wissenschaftliche Güte des Instruments als summative Evaluationsmethode für mentale Beanspruchung auf Basis der in Teilabschnitt 3.2.5 vorgestellten nicht formalen Gütekriterien hinweisen (insbesondere Sensititivität und Diagnostizität), jedoch konnte auch bereits gezeigt werden, dass thermische Aspekte in der Umwelt dabei einen großen Einfluss auf die Vasokonstriktion haben und die Messung leicht verzerren können (S. Miyake et al., 2009). In Kombination mit den fehlenden formalen empirischen Nachweisen bezüglich der psychometrischen Gütekriterien eignet sich die PPG auch nicht für eine objektiv summative Evaluation intuitiver Benutzung im Rahmen des Projekts 3D-GUIde. Die PPG qualifiziert sich daher auch nicht als Quasi-Außenkriterium für eine Meta-Evaluation einer summativen Evaluationsmethode für intuitive Benutzung.

Aktuelle Reviews zur objektiven Messung mentaler Beanspruchung mit physiologischen Maßen (z.B. Charles & Nixon, 2019; Cowley et al., 2016) enthalten ferner keine empirischen Nachweise bezüglich der zeitlichen Anwendungseffizienz von Maßen der Herz- und Kreislaufaktivität im Vergleich zu anderen objektiven Methoden, weswegen den hier vorgestellten Maßen keine hohe zeitliche Anwendungseffizienz attestiert werden kann.

Maße der Gehirnaktivität

Neben den eben vorgestellten Maßen der Herz- und Kreislauftätigkeit existieren auch eine Reihe von Maßen, um die mentale Beanspruchung anhand der Gehirnaktivität beurteilen zu können. Mithilfe eines Elektroenzephalogramms (EEG) können zwei Arten von Indikatoren aufgezeichnet werden, die am häufigsten zur Abbildung mentaler Beanspruchung eingesetzt werden und aus der Perspektive von einer Reihe von Wissenschaftlern eine hohe Inhaltsvalidität besitzen (z.B. Brookhuis & De Waard, 1993, 2010; Brunken, Plass, & Leut-

ner, 2003; Charles & Nixon, 2019; F. Chen et al., 2016; Cowley et al., 2016; G. F. Wilson & Russell, 2003; Young et al., 2015): *Frequenzbänder* und *ereigniskorrelierte Potentiale*.

Wie das EKG kann das EEG auch durch eine Analyse der spektralen Leistungsdichte, was üblicherweise mithilfe einer Fourier-Transformation durchgeführt wird, in mehrere Frequenzbänder unterteilt werden (Kramer, 1991), welche üblicherweise zwischen 1 und 40 Hertz liegen und als Delta (bis zu 2 Hertz), Theta (4 bis 7 Hertz), Alpha (8 bis 14 Hertz) und Beta (14 bis 25 Hertz) bezeichnet werden (Charles & Nixon, 2019). Eine erhöhte mentale Beanspruchung lässt sich anhand eines Rückgangs der Aktivität im Alpha-Band und der gleichzeitigen Steigerung der Aktivität im Theta-Band feststellen (Kramer, 1991).

Die wissenschaftliche Güte von Frequenzbändern als objektives Maß für mentale Beanspruchung wurde in den von Cowley et al. (2016) und Charles und Nixon (2019) genannten Arbeiten überwiegend durch die in Teilabschnitt 3.2.5 vorgestellten nicht formalen Gütekriterien (insbesondere Sensitivität und Diagnostizität) sichergestellt, die für einen vollständigen Nachweis der wissenschaftlichen Güte nicht ausreichend sind (Hancock & Matthews, 2019; Matthews et al., 2015). Dementsprechend attestieren zwar eine Reihe von Forschern in ihren Arbeiten Frequenzbändern auf Basis dieser nicht formalen Kriterien ausreichende wissenschaftliche Güte (z.B. Brookings, Wilson, & Swain, 1996; Fournier, Wilson, & Swain, 1999; Klimesch, 1997; Ryu & Myung, 2005), wohingegen andere Forscher diese anzweifeln. Als Gründe sind hier beispielsweise eine starke Anfälligkeit von Frequenzbändern gegenüber motorischen Bewegungen (z.B. Fournier et al., 1999), eine Einschränkung der maximalen Messdauer (z.B. Fairclough, Venables, & Tattersall, 2005), eine mangelnde Unterscheidbarkeit von feineren Abstufungen (d.h. lediglich Unterscheidung zwischen aktiven Phasen und Ruhephasen) bei der Messung mentaler Beanspruchung (z.B. Hankins & Wilson, 1998) und die Anfälligkeit gegenüber der eigenen Atemfrequenz (z.B. Zhang, Yu, & Xie, 2010) zu nennen. Da neben diesen Limitationen auch ein formaler, empirischer Nachweis der wissenschaftlichen Güte bezüglich der psychometrischen Gütekriterien fehlt, kommen Frequenzbänder nicht für eine objektiv summative Evaluation intuitiver Benutzung im Rahmen des Projekts 3D-GUIde in Frage. Sie eignen sich daher auch nicht als Quasi-Außenkriterium bei einer Meta-Evaluation einer summativen Evaluationsmethode für intuitive Benutzung.

Aufgrund dieser Einschränkungen kommen anstelle von Frequenzbändern, auch häufig ereigniskorrelierte Potentiale zur Abbildung mentaler Beanspruchung zum Einsatz. Diese können im Gegensatz zu Frequenzbändern die mentale Beanspruchung im Zeitverlauf widerspiegeln. Ereigniskorrelierte Potentiale sind zeitlich an einen bestimmten Stimulus gebunden, wie z.B. an den Beginn eines Geräusches oder die Präsentation eines Bildes. Sie erlauben dem Evaluator auf diese Weise sich ein Bild von der mentalen Beanspruchung während der Handlungsregulation zu machen. Mehrere ereigniskorrelierte Potentiale mit hoher Latenzzeit (d.h. positive oder negative Potentiale, die nach 100, 200 oder 300 Millisekunden nach der Präsentation des Stimulus auftauchen) werden dabei als Indikatoren mentaler Beanspruchung berücksichtigt (Cowley et al., 2016), wohingegen ereigniskorrelierte Potentiale niedriger Latenzen keinen Zusammenhang mit mentaler Beanspruchung zeigen (Nittono, Hamada, & Hori, 2003). Das Ansteigen der Latenzen bei einigen Komponenten (Ying, Fu, Qian, & Sun, 2011) und der Abfall bei anderen (M. W. Miller, Rietschel, McDonald, & Hatfield, 2011) zeigt eine Änderung in der mentalen Beanspruchung eines

Menschen an. Dabei kommt die Amplitude der P300-Komponente (d.h. Amplitude tritt mit einer Latenz von 300 bis $\leq$ 700 Millisekunden zum auslösenden Reiz auf) laut der aktuellen Reviews von Cowley et al. (2016) und Charles und Nixon (2019) am häufigsten für die summative Evaluation mentaler Beanspruchung zum Einsatz. Die wissenschaftliche Güte der P300-Komponente als objektives Maß für mentale Beanspruchung wurde in den von Cowley et al. (2016) und Charles und Nixon (2019) genannten Arbeiten überwiegend durch die in Teilabschnitt 3.2.5 vorgestellten nicht formalen Gütekriterien (insbesondere Sensitivität und Diagnostizität) sichergestellt, die für einen vollständigen Nachweis der wissenschaftlichen Güte nicht ausreichend sind (Hancock & Matthews, 2019; Matthews et al., 2015). Insbesondere eine durch einen auditiven Stimulus ausgelöste P300 spiegelt die mentale Beanspruchung eines Handelnden sehr gut wider (Cowley et al., 2016; Kok, 2001).

Um die mentale Beanspruchung mithilfe von ereigniskorrelierten Potentialen in der HCI überhaupt messen zu können, muss der Handelnde entweder durch einen externen Stimulus (z.B. Geräusch, Lichtreiz) leicht abgelenkt werden (z.B. Hohnsbein, Falkenstein, & Hoormann, 1995) oder der Handelnde wird bei der Handlung mit einer kontinuierlichen Abfolge von Stimuli (z.B. Nutzen eines Musikstücks zur Stimulation, siehe Poikonen et al., 2016) konfrontiert oder das Verhalten des Nutzers im Rahmen der Handlungsregulation (z.B. Nutzen der Mausklicks des Nutzers, siehe Nittono et al., 2003) verursacht selbst geeignete Stimuli für eine P300-Messung (Cowley et al., 2016). An dieser Stelle soll angemerkt werden, dass sich aufgrund dieser Einschränkungen nicht alle Arten von Handlungen für eine Evaluation mithilfe von ereigniskorrelierten Potentialen eignen und das Nutzen von externen Stimuli den Nutzer leicht ablenken, verärgern oder erschöpfen kann, was wiederum die wissenschaftliche Güte der Methode stark gefährdet (Cowley et al., 2016; Kramer & Spinks, 1991; Trejo, Kramer, & Arnold, 1995; Ullsperger & Von Cramon, 2001). Aufgrund dieser Limitationen und des fehlenden formalen empirischen Nachweises bezüglich der psychometrischen Gütekriterien kommen auch ereigniskorrelierte Potentiale nicht für eine objektiv summative Evaluation intuitiver Benutzung im Rahmen des Projekts 3D-GUIde in Frage. Sie eignen sich daher auch nicht als Quasi-Außenkriterium für eine Meta-Evaluation einer summativen Evaluationsmethode für intuitive Benutzung.

Aktuelle Reviews zur objektiven Messung mentaler Beanspruchung mit physiologischen Maßen (z.B. Charles & Nixon, 2019; Cowley et al., 2016) enthalten ferner keine empirischen Nachweise bezüglich der zeitlichen Anwendungseffizienz von Maßen der Gehirnaktivität im Vergleich zu anderen objektiven Methoden, weswegen den hier vorgestellten Maßen keine hohe zeitliche Anwendungseffizienz attestiert werden kann.

Maße der elektrodermalen Aktivität

Die mentale Beanspruchung lässt sich auch über die Haut eines Menschen objektiv erfassen. Über eine Messung der elektrodermalen Aktivität (EDA) können autonome Änderungen im elektrischen Leitungswiderstand der Haut (d.h. durch erhöhte Schweißsekretion) erkannt werden. Diese Änderungen werden durch eine Erregung des sympathischen Nervensystems ausgelöst (Dawson, Schell, & Filion, 2017). Eine Reihe von empirischen Arbeiten konnten bereits zeigen, dass der elektrische Leitungswiderstand der Haut positiv mit mentaler Beanspruchung korreliert (z.B. Charles & Nixon, 2019; F. Chen et al., 2016; Cowley

et al., 2016; Mehler, Reimer, Coughlin, & Dusek, 2009; Nourbakhsh, Wang, & Chen, 2013; Nourbakhsh, Wang, Chen, & Calvo, 2012; Paxion et al., 2014; Shi, Ruiz, Taib, Choi, & Chen, 2007). Die Messung der EGA ist sehr kostengünstig und kann einfach durchgeführt werden, da dafür lediglich ein bis zwei Sensoren auf der Hand oder dem Fuß angebracht werden müssen (F. Chen et al., 2016).

Die wissenschaftliche Güte der EDA als objektives Maß für mentale Beanspruchung wurde in den von Cowley et al. (2016) und Charles und Nixon (2019) genannten Arbeiten überwiegend durch die in Teilabschnitt 3.2.5 vorgestellten nicht formalen Gütekriterien sichergestellt, welche für einen vollständigen Nachweis der wissenschaftlichen Güte nicht ausreichend sind (Hancock & Matthews, 2019; Matthews et al., 2015). Dementsprechend attestieren eine Reihe von Forschern in ihren Arbeiten der EDA auf Basis dieser nicht formalen Kriterien (insbesondere Sensitivität und Diagnostizität) wissenschaftliche Güte (z.B. Collet, Clarion, Morel, Chapon, & Petit, 2009; Fairclough & Venables, 2006; Mehler et al., 2009; Shi et al., 2007; G. F. Wilson, 2002), wohingegen andere Forscher diese anzweifeln. Als Gründe sind hier beispielsweise eine starke Anfälligkeit der Hautleitfähigkeit für Stress und eine Verzerrung durch den allgemeinen Erregungsgrad (z.B. Ikehara & Crosby, 2005), eine mangelnde Unterscheidbarkeit von feineren Abstufungen (d.h. lediglich Unterscheidung zwischen aktiven Phasen und Ruhephasen) bei der Messung mentaler Beanspruchung (z.B. Haapalainen, Kim, Forlizzi, & Dey, 2010; Widyanti, 2017) und die starke Beeinflussbarkeit durch Temperatur, Feuchtigkeit, das Geschlecht des Nutzers, seines Alters, sowie der Tages- und Jahreszeit zu nennen. Eine derartige Anfälligkeit macht eine Messung dieses Merkmals mit hoher wissenschaftlicher Güte sehr schwierig (z.B. Kramer, 1991). Aufgrund dieser Limitationen und des fehlenden formalen empirischen Nachweises bezüglich der psychometrischen Gütekriterien kommt auch die elektrodermale Aktivität nicht für eine objektiv summative Evaluation intuitiver Benutzung im Rahmen des Projekts 3D-GUIde in Frage. Sie eignet sich daher auch nicht als Quasi-Außenkriterium für eine Meta-Evaluation einer summativen Evaluationsmethode für intuitive Benutzung.

Aktuelle Reviews zur objektiven Messung mentaler Beanspruchung mit physiologischen Maßen (z.B. Charles & Nixon, 2019; Cowley et al., 2016) enthalten ferner keine empirischen Nachweise bezüglich der zeitlichen Anwendungseffizienz von Maßen der elektrodermalen Aktivität im Vergleich zu anderen objektiven Methoden, weswegen den hier vorgestellten Maßen keine hohe zeitliche Anwendungseffizienz attestiert werden kann.

Maße der respiratorischen Aktivität

Die mentale Beanspruchung lässt sich außerdem noch über die Atmung messen (Charles & Nixon, 2019; F. Chen et al., 2016; Cowley et al., 2016; L. J. Mulder, 1992; Roscoe, 1992). Dabei kann mentale Beanspruchung durch die Atemfrequenz, den Luftstrom, das Atemvolumen und die Atemgasanalyse erfasst werden (Charles & Nixon, 2019; Cowley et al., 2016). Die *Atemfrequenz* stellt laut Roscoe (1992) von allen eben genannten Maßen das nützlichste Maß dar, da sie leichter (d.h. effizienter und weniger invasiv) als die anderen Maße zu messen ist (z.B. Messung der Brustumfangsveränderung durch einen Atemgürtel vs. Atemgasanalyse mithilfe einer Maske über Nase und Mund). Die Atemfrequenz liegt bei erhöhter mentaler Beanspruchung höher (Charles & Nixon, 2019; F. Chen et al., 2016;

Cowley et al., 2016; Grassmann, Vlemincx, von Leupoldt, Mittelstädt, & Van den Bergh, 2016; Roscoe, 1992).

Die wissenschaftliche Güte der Atemfrequenz als objektives Maß für mentale Beanspruchung wurde in den von Cowley et al. (2016) und Charles und Nixon (2019) genannten Arbeiten überwiegend durch die in Teilabschnitt 3.2.5 vorgestellten nicht formalen Gütekriterien sichergestellt, die für einen vollständigen Nachweis der wissenschaftlichen Güte nicht ausreichend sind (Hancock & Matthews, 2019; Matthews et al., 2015). Dementsprechend attestieren eine Reihe von Forschern in ihren Arbeiten der Atemfrequenz auf Basis dieser nicht formalen Kriterien (insbesondere Sensitivität und Diagnostizität) wissenschaftliche Güte (z.B. Backs, Navidzadeh, & Xu, 2000; Backs & Seljos, 1994; Brooke, 1996; Fairclough & Venables, 2006; Veltman & Gaillard, 1998; Zhang et al., 2010), wohingegen andere Forscher diese anzweifeln. Als Gründe sind hier beispielsweise eine Einschränkung der maximalen Messdauer (z.B. Fournier et al., 1999), eine Anfälligkeit gegenüber der physischen Aktivität und des eigenen Metabolismus (z.B. Cowley et al., 2016; Grassmann et al., 2016) und die Tatsache zu nennen, dass Sprechen die Atmung unterbrechen und dadurch Atemmuster verändern kann, was alles zu Änderungen in der Atemfrequenz führt, die nicht durch erhöhte mentale Beanspruchung verursacht wurden (z.B. Bernardi et al., 2000; Roscoe, 1992; Sirevaag et al., 1993). Aufgrund dieser Limitationen und des fehlenden formalen empirischen Nachweises bezüglich der psychometrischen Gütekriterien kommt auch die Atemfrequenz nicht für eine objektiv summative Evaluation intuitiver Benutzung im Rahmen des Projekts 3D-GUIde in Frage. Sie eignet sich daher auch nicht als Quasi-Außenkriterium für eine Meta-Evaluation einer summativen Evaluationsmethode für intuitive Benutzung.

Aktuelle Reviews zur objektiven Messung mentaler Beanspruchung mit physiologischen Maßen (z.B. Charles & Nixon, 2019; Cowley et al., 2016) enthalten ferner keine empirischen Nachweise bezüglich der zeitlichen Anwendungseffizienz von Maßen der respiratorischen Aktivität im Vergleich zu anderen objektiven Methoden, weswegen den hier vorgestellten Maßen keine hohe zeitliche Anwendungseffizienz attestiert werden kann.

Maße der okulomotorischen Aktivität

Als letztes verbleibt für die summative Evaluation intuitiver Benutzung anhand von mentaler Beanspruchung geeignete Biosignal noch die Augenaktivität, welche mithilfe von Elektrookulografie, Eye Trackern oder einfachen Infrarotkameras aufgezeichnet werden kann (siehe Charles & Nixon, 2019; F. Chen et al., 2016; Siyuan Chen & Epps, 2013; Cowley et al., 2016; Lipp & Neumann, 2004; Schneider, 2019). In den letzten Jahren wurde entsprechende Ausrüstung immer günstiger und somit zugänglicher, weswegen dieses Maß immer häufiger angewendet wird (Charles & Nixon, 2019). Wie schon bei den anderen bereits vorgestellten physiologischen Maßen, können bei der Verwendung von Blickdaten verschiedene Parameter zur Beurteilung von mentaler Beanspruchung herangezogen werden. Okulare Parameter können laut Rötting (2001) in Augen- und Blickbewegungen unterschieden werden. Augenbewegungen können dabei alleinig durch die Bewegung des Auges erfasst werden, wohingegen Blickbewegungen ohne einen Bezug zum konkreten Objekt, dass gerade beobachtet wird, nicht interpretierbar sind. Die beiden gängigsten Parameter innerhalb der Gruppe der Augenbewegungen sind *Sakkadenlängen* (SL) und

Fixationsdauern (FD). Aufgrund der Tatsache, dass das menschliche Blickfeld eine übliche Ausprägung von etwa 100 Grad und der foveale Bereich des scharfen Sehens sogar noch kleiner ausfällt, sind Augenbewegungen zur Wahrnehmung von Objekten notwendig (Rötting, 2001).

Bei Sakkaden handelt es sich um schnelle, ruckartige Bewegungen des Auges, um den Blick auf neue Objekte richten zu können, ohne dass bereits eine Informationsaufnahme stattfindet. Sie können sowohl ungewollt durch Bewegungsänderungen im peripheren Sichtfeld ausgelöst werden als auch gewollt bei der Wahrnehmung und Interpretation einer neuen Informationsquelle (Rötting, 2001). Eine Reihe von theoretischen und empirischen Arbeiten konnten die Sensitivität der Sakkadenlänge mit Such-, Informationsverarbeitungs- und Wahrnehmungsprozessen in Verbindung bringen (z.B. Matthews et al., 2015; May, Kennedy, Williams, Dunlap, & Brannan, 1990; Paxion et al., 2014; Rötting, 2001). Ausgehend von diesen Arbeiten wird eine durchschnittliche Sakkadenlänge von 30 Millisekunden als Indikator für erhöhte mentale Beanspruchung gewertet. Neben dem Anstieg der Dauer von sakkadischen Augenbewegungen indiziert auch ein massiver Einbruch der Sakkadengeschwindigkeit (z.B. Sakkaden können Geschwindigkeiten bis 1000 Grad pro Sekunde erreichen) eine erhöhte mentale Beanspruchung. Di Stasi et al. (2010) analysierten dazu Geschwindigkeitsspitzen und konnten bei erhöhter mentaler Beanspruchung einen starken Abfall in der Geschwindigkeit (d.h. 7.2 Grad pro Sekunde) feststellen. Darüber hinaus konnten Bodala, Ke, Mir, Thakor und Al-Nashash (2014) Sakkadengeschwindigkeitsspitzen als Maß für mentale Beanspruchung etablieren.

Nimmt das Auge eine Ruheposition ein, spricht man von einer Fixation (Rötting, 2001). Im Gegensatz zu Sakkaden können mithilfe von Fixationen Informationen aufgenommen werden, wofür laut Sträter (2016) minimal 175 Millisekunden benötigt werden, was sich aber individuell aufgrund diverser Einflussfaktoren (z.B. Umwelt, Alter) unterscheiden kann. Eine Reihe von empirischen Arbeiten legen nahe, dass die Fixaktionsdauer die benötigte Zeit der Informationsaufnahme und auf diese Weise die damit verbundene mentale Beanspruchung bei der kognitiven Informationsverarbeitung abbildet (z.B. Dehais, Causse, Vachon, & Tremblay, 2012; Matthews et al., 2015; Meyberg, Werkle-Bergner, Sommer, & Dimigen, 2015; Rosch & Vogel-Walcutt, 2013; Tsai, Viirre, Strychacz, Chase, & Jung, 2007). Die wissenschaftliche Güte von Sakkaden und Fixationen als objektive Maße für mentale Beanspruchung wurde in den von Cowley et al. (2016) und Charles und Nixon (2019) genannten Arbeiten überwiegend durch die in Teilabschnitt 3.2.5 vorgestellten nicht formalen Gütekriterien (insbesondere Sensitivität und Diagnostizität) sichergestellt, die für einen vollständigen Nachweis der wissenschaftlichen Güte nicht ausreichend sind (Hancock & Matthews, 2019; Matthews et al., 2015). In den hier genannten Arbeiten attestierten die Forscher Sakkaden und Fixationen auf Basis solcher nicht formalen Kriterien eine ausreichende wissenschaftliche Güte. Kritische Befunde werden zu diesem Thema in den Reviews von Cowley et al. (2016) und Charles und Nixon (2019) nicht erwähnt, da das Maß auch nur am Rande besprochen wurde. Aufgrund des fehlenden formalen empirischen Nachweises bezüglich der psychometrischen Gütekriterien kommen jedoch auch Sakkadenlängen und Fixationsdauern nicht für eine objektiv summative Evaluation intuitiver Benutzung im Rahmen des Projekts 3D-GUIde in Frage. Sie eignen sich daher auch nicht als Quasi-Außenkriterien für eine Meta-Evaluation einer summativen Evaluationsmethode für intuitive Benutzung.

Neben Sakkadenlängen und Fixationsdauern kann das Blickverhalten auch komplexer abgebildet werden. Es existieren beispielsweise Arbeiten, in denen die mentale Beanspruchung anhand des *Nearest Neighbor Index (NNI)* bestimmt wurde. Basierend auf der Forschungsarbeit von Harris Sr, Tole, Ephrath und Stephens (1982) kann von einem Zusammenhang zwischen mentaler Beanspruchung und der Entropie (d.h. Maß der Unordnung) des Blickverhaltens ausgegangen werden. Eine Reihe von empirischen Arbeiten konnte zeigen, dass der NNI sehr sensibel auf Änderungen in der mentalen Beanspruchung reagiert, sich aber auch proportional zu einer zunehmenden Entropie verhält (Di Nocera, Camilli, & Terenzi, 2007; Di Nocera, Terenzi, & Camilli, 2006; Schneider, 2019). Der NNI wird als das Verhältnis zwischen dem durchschnittlichen Mindestabstand von Fixationen in einer vorliegenden Verteilung und dem mittleren Abstand einer vollständig zufälligen bzw. regellosen Verteilung von der gleichen Anzahl an Fixationen verstanden (Di Nocera et al., 2006). Das Problem bei der Berechnung des NNI ist die Definition der Bezugsfläche, auf der die Fixationen stattfinden. Befindet sich beispielsweise eine Fixation außerhalb des aufgabenrelevanten Blickfeldes, weil der Nutzer bei der Aufgabenbearbeitung von seiner Umwelt abgelenkt wurde, verzerrt dies sofort den NNI, was seine wissenschaftliche Güte fragwürdig macht. Zur Sicherstellung der wissenschaftlichen Güte des NNI als Maß mentaler Beanspruchung kamen nur die in Teilabschnitt 3.2.5 genannten nicht formalen Kriterien (insbesondere Sensitivität und Diagnostizität) zum Einsatz (Di Nocera et al., 2007; Di Nocera et al., 2006; Schneider, 2019). Der NNI wurde darüber hinaus von aktuellen Reviews zu diesem Thema bisher noch nicht thematisiert (siehe Charles & Nixon, 2019; Cowley et al., 2016). Aufgrund des fehlenden formalen empirischen Nachweises bezüglich der psychometrischen Gütekriterien kommt auch der NNI nicht für eine objektiv summative Evaluation intuitiver Benutzung im Rahmen des Projekts 3D-GUIde in Frage. Er eignet sich daher auch nicht als Quasi-Außenkriterium für eine Meta-Evaluation einer summativen Evaluationsmethode für intuitive Benutzung.

Neben den vorgestellten, eher bewusst beeinflussbaren okularen Parametern beschäftigt sich die Forschungsliteratur zunehmend auch mit einer Reihe weiterer Parameter (z.B. Blinzelrate, Blinzeldauer, Pupillendurchschnittsvariabilität), die sich weniger stark bewusst beeinflussen lassen (Charles & Nixon, 2019). Die Blinzelrate wird dabei als die Anzahl an Lidschlüssen (d.h. ein Lidschluss dauert in der Regel zwischen 70 und 100 Millisekunden) in einem bestimmten Zeitintervall definiert (Schneider, 2019). Eine Reihe von empirischen Untersuchungen konnten bereits zeigen (z.B. Siyuan Chen & Epps, 2014; Faure, Lobjois, & Benguigui, 2016; Recarte, Pérez, Conchillo, & Nunes, 2008; Veltman & Gaillard, 1996), dass ein Zusammenhang zwischen mentaler Beanspruchung und der Blinzelrate existiert. Eine Zunahme der Blinzelrate bedeutet dabei eine steigende mentale Beanspruchung (Schneider, 2019). Die Blinzeldauer, welche zwischen 70 und 500 Millisekunden liegen kann, gilt dabei jedoch als zuverlässigeres Merkmal als die Blinzelrate (Schneider, 2019). Diese nimmt mit zunehmender mentaler Beanspruchung ab. Die Blinzeldauer kann entweder als durchschnittliche oder kumulierte Blinzeldauer angegeben werden. Die wissenschaftliche Güte der Blinzelrate und Blinzeldauer als objektive Maße für mentale Beanspruchung wurde in den von Cowley et al. (2016) und Charles und Nixon (2019) genannten Arbeiten überwiegend durch die in Teilabschnitt 3.2.5 vorgestellten nicht formalen Gütekriterien (insbesondere Sensitivität und Diagnostizität) sichergestellt, die für einen vollständigen Nachweis der wissenschaftlichen Güte nicht ausreichend sind (Hancock & Matthews, 2019; Matthews et al., 2015). Dementsprechend attestieren eine

Reihe von Forschern in ihren Arbeiten den beiden Maßen auf Basis dieser nicht formalen Kriterien ausreichende wissenschaftliche Güte (z.B. Brooke, 1996; De Rivecourt et al., 2008; Fairclough & Venables, 2006; Ryu & Myung, 2005; Sirevaag et al., 1993; Veltman & Gaillard, 1996), wohingegen andere Forscher diese anzweifeln. So zeigten Fairclough et al. (2005), dass Änderungen in der Blinzelrate und der Blinzeldauer nur kurzfristig für die Evaluation mentaler Beanspruchung geeignet sind, da sie nur in der ersten Hälfte ihres Experiments (d.h. ersten 32 Minuten) funktionierten. Aufgrund dieser Limitationen und des fehlenden formalen empirischen Nachweises bezüglich der psychometrischen Gütekriterien kommen auch Blinzelrate und Blinzeldauer nicht für eine objektiv summative Evaluation intuitiver Benutzung im Rahmen des Projekts 3D-GUIde in Frage. Sie eignen sich daher auch nicht als Quasi-Außenkriterien für eine Meta-Evaluation einer summativen Evaluationsmethode für intuitive Benutzung.

Als letzter physiologischer okularer Parameter soll nun die Pupillendurchschnittsvariabilität (PDV) betrachtet werden. Mithilfe der Pupille, welche eine Öffnung in der Iris darstellt (d.h. kann sich zwischen 2 und 8 Millimetern öffnen), kann das menschliche Auge die Menge des dort einfallenden Lichts regulieren, brechen und bündeln (Charles & Nixon, 2019; Cowley et al., 2016; Schneider, 2019). Veränderungen der Pupillengröße können dabei von verschiedenen Prozessen verursacht werden. Beispielsweise reguliert der *pupillometrische Lichtreflex* die Menge des einfallenden Lichts und adaptiert die Pupille an vorliegende Beleuchtungsbedingungen (d.h. in hellen Umgebungen ist die Pupille klein und dunklen Umgebungen groß) (Schneider, 2019). Ein weiterer Reflex ist der sogenannte *Akkomodationsreflex*, welche die Pupillenkrümmung verändert, was ebenfalls Auswirkungen auf die Pupillengröße hat. Dieser Reflex ist entscheidend, um mithilfe der Krümmung Objekte in unterschiedlichen Entfernungen scharf sehen zu können (Schneider, 2019). Neben diesen rein optischen Reflexen existieren bestimmte Adaptionen, mit deren Hilfe mentale Aktivitäten und die daraus resultierende mentale Beanspruchung abgeleitet werden kann (Schneider, 2019). Die Interpretation von aufgabenkorrelierten Reflexen in Bezug auf mentale Beanspruchung setzt dabei immer voraus, dass die eben beschriebenen reflexartigen Pupillenveränderungen, die von Helligkeitsunterschieden und Akkommodationsprozessen verursacht werden, ausgeschlossen werden können (Schneider, 2019).

Der *psychosensorische Reflex* gilt als Indikator für Vorgänge innerhalb des zentralen Nervensystems. Er indiziert auf diese Weise die Dynamik und Intensität der menschlichen Informationsverarbeitung der damit verbundenen mentalen Beanspruchung (Schneider, 2019). Bei diesem Reflex handelt sich um unregelmäßige, sehr abrupte Fluktuationen der Pupille, welche häufig durch extrem schnelle Anstiege gefolgt von ebenso schnellen Abstiegen in der Pupillengröße charakterisiert sind (Schneider, 2019). Die Pupille weitet sich dabei solange kontinuierlich, bis sie nach 700 bis 1200 Millisekunden ihre Amplitude erreicht hat und kehrt anschließend wieder schnell auf ihr Ausgangsniveau zurück, sofern der Reiz verschwunden ist (Manzey, 1997). Die Amplitudenhöhe dieser phasischen Pupillenreaktion korreliert dann mit der mentalen Beanspruchung und quantifiziert diese. Die wissenschaftliche Güte des psychosensorischen Reflexes als objektives Maß für mentale Beanspruchung wurde in den von Cowley et al. (2016) und Charles und Nixon (2019) genannten Arbeiten überwiegend durch die in Teilabschnitt 3.2.5 vorgestellten nicht formalen Gütekriterien sichergestellt (insbesondere Sensitivität und Diagnostizität), die für einen vollständigen Nachweis der wissenschaftlichen Güte nicht ausreichend sind (Hancock & Matthews, 2019; Matthews et al., 2015). Dementsprechend attestieren eine Reihe von

Forschern in ihren Arbeiten auf Basis dieser nicht formalen Kriterien dem psychometrischen Reflex eine ausreichend wissenschaftliche Güte (z.B. Causse et al., 2010; Gao, Wang, Song, Li, & Dong, 2013; Just & Carpenter, 1993; Recarte et al., 2008), wohingegen andere Forscher diese anzweifeln. Als Gründe sind hier beispielsweise eine starke Abhängigkeit des psychometrischen Reflexes von der Umgebungsbeleuchtung (z.B. De Rivecourt et al., 2008) und die Tatsache zu nennen, dass mithilfe des Maßes auf Basis einiger empirischer Arbeiten nur grobe Unterschiede in der mentalen Beanspruchung (d.h. Unterscheidung zwischen aktiven Phasen und Ruhephasen) festgestellt werden können (z.B. Schultheis & Jameson, 2004). Aufgrund dieser Limitationen und des fehlenden formalen empirischen Nachweises bezüglich der psychometrischen Gütekriterien kommt auch der psychosensorische Reflex nicht für die objektiv summative Evaluation intuitiver Benutzung im Rahmen des Projekts 3D-GUIde in Frage. Er eignet sich auch nicht als Quasi-Außenkriterium für eine Meta-Evaluation einer summativen Evaluationsmethode intuitiver Benutzung.

Alle hier vorgestellten okularen Parameter werden generell im Labor erhoben und sind deswegen in der Praxis beispielsweise während eines Nutzertests nur begrenzt anwendbar, da sie gleichbleibende Beleuchtungsverhältnisse während der gesamten Versuchssituation voraussetzen (Schneider, 2019). Wie bereits erwähnt kann laut De Rivecourt et al. (2008) eine Veränderung des Pupillendurchmessers auch bereits durch eine Änderung in der Umgebungsbeleuchtung ausgelöst werden und muss nicht unbedingt das Resultat erhöhter mentaler Beanspruchung sein. Darüber hinaus berücksichtigen die meisten Arbeiten individuelle Unterschiede in der Pupillengröße unzureichend (Schneider, 2019). Die wissenschaftliche Güte okularer Parameter wird auf diese Weise herabgesetzt. Ferner konnten Wiebelitz und Schmitz (1983) zeigen, dass die Frequenz des Lichts bei okularen Parametern ein hohes Störpotential bietet, da die Fähigkeit der Pupillenreaktion direkt von der Frequenz des einfallenden Lichts beeinflusst wird. Besonders erheblich ist dieser Einfluss im Frequenzbereich von 25 bis 150 Hertz, was der Bildwiederholungsrate bei Monitoren, Smartphones und anderen HCI-relevanten Medien entspricht (siehe Schneider, 2019). Darüber hinaus konnten Wiebelitz und Schmitz (1983) zeigen, dass auch die Außenbeleuchtung sofern diese künstlich ist (z.B. Neonröhren im Büro) einen Einfluss auf die Pupillenweite haben kann. Des Weiteren haben unterschiedliche Farben nachgewiesenermaßen Einfluss auf die Pupillenbewegung (Wiebelitz & Schmitz, 1983). Bouma (1962) konnte zeigen, dass die Wellenlänge des Lichts Auswirkung auf den Pupillendurchmesser hat. Laut Galley (2001) ist dafür das subjektive Lichtempfinden ausschlaggebend, da die Retina Unterschiede bei der Lichtempfindlichkeit unter Bestrahlung verschiedener Farben aufweist. Dieser Effekt lässt sich dadurch kontrollieren, dass man die Leuchtdichte (d.h. Flächenhelligkeit, mit welcher das Auge die Fläche wahrnimmt) der vorhandenen Farben konstant hält (Ishigaki, Miyao, & Ishihara, 1991).

Bei der Evaluation eines Systems im Rahmen eines menschzentrierten Gestaltungsprozesses ist dies aber meist nicht möglich, da die Stimuli (z.B. zu testende Webseiten) ja dynamisch sind. Diese Einflussfaktoren so zu kontrollieren, sodass noch wissenschaftliche valide Ergebnisse damit erzielt werden können, stellt demnach eine hohe Herausforderung bei der Verwendung von okularen Parametern dar. Da die Pupille während eines Nutzertests üblicherweise mit einem Eye Tracker bzw. einer Infrarotkamera erfasst wird, muss bei der Ableitung des Pupillendurchmessers auf jeden Fall auch die Eigenstrahlung des Ausgabegerätes (z.B. Monitor) berücksichtigt werden. Die Analyse der Pupille wird also durch die Kontrollierbarkeit derartiger Störreize bedingt. Aufgrund der hohen Lichtempfindlich-

keit von Pupillenmessungen, schlugen Pfleging, Fekety, Schmidt und Kun (2016) daher ein Modell zur Klassifikation von mentaler Beanspruchung anhand des Pupillendurchmessers bei verschiedenen Lichtverhältnissen vor. Ein anderer Ansatz, der in aktuellen Arbeiten (z.B. Cowley et al., 2016; Schneider, 2019) stark diskutiert wird, stammt von S. P. Marshall (2002), bei dem nicht mehr die ganze Pupillenfläche analysiert wird, sondern nur der sogenannte *Index of Cognitive Activity (ICA)* berechnet wird. Dieser versucht mithilfe einer Wavelet-Analyse die psychosensorischen Reflexe erfolgreich von den Licht- und Akkomodationsreflexen zu trennen. Eine Wavelet-Analyse ist in der Verarbeitung und Dekomposition von Signalen ein weit verbreitetes Verfahren, mit dem ein Signal orthogonal transformiert wird, sodass die Frequenz über die Zeit aufgetragen werden kann (Schneider, 2019). Eine Reihe von empirischen Arbeiten konnten bereits die Sensitivität des ICA hinsichtlich mentaler Beanspruchung unabhängig von Beleuchtungsbedingungen bei diversen Laboruntersuchungen demonstrieren (z.B. Demberg, 2013; P. Marshall, Rogers, & Hornecker, 2007; S. P. Marshall, 2002; S. P. Marshall, Pleydell-Pearce, & Dickson, 2003; Schwalm, 2009). Leider konnten diese Arbeiten noch nicht die wissenschaftliche Güte des ICA bezüglich der formalen psychometrischen Gütekriterien bestätigen, weswegen der ICA nicht für eine objektiv summative Evaluation intuitiver Benutzung im Rahmen des Projekts 3D-GUIde in Frage kommt. Er eignet sich daher auch nicht als Quasi-Außenkriterium für eine Meta-Evaluation einer summativen Evaluationsmethode für intuitive Benutzung.

Neben der Lichtempfindlichkeit existieren noch eine Reihe weiterer Störvariablen, die einen Einfluss auf die Pupillometrie haben können (Kerkau, 2006). So wird der Pupillendurchmesser mit zunehmendem Alter größer, was bei der Messung durch das Erheben einer individuellen Baseline berücksichtigt werden muss (Kerkau, 2006). Des Weiteren können auch Medikamente, Drogen und Alkohol Einfluss auf die Pupille haben (Charles & Nixon, 2019). Hier ist insbesondere die Arbeit von Gambill, Ogle und Kearns (1967) aus dem medizinischen Bereich zu nennen, die zeigt dass Pupillometrie für Personen, die unter dem Einfluss bestimmter Medikamente stehen ungeeignet ist. Die reine körperliche Anstrengung korreliert ebenfalls mit dem Pupillendurchmesser (Ishigaki et al., 1991), stellt aber während eines klassischen Nutzertests meistens kein Problem dar. Neben diesen Störeinflüssen sollte ferner darauf hingewiesen werden, dass in den meisten hier referenzierten Studien sehr kurze Stimuli verwendet wurden, was für eine realistische Anwendung zur summativen Evaluation intuitiver Benutzung im Rahmen eines menschzentrierten Gestaltungsprozesses als nicht ausreichend erscheint. Eine Evaluation in realen, komplexen und weniger kontrollierten Umgebungen ist dementsprechend zwingend notwendig, um okulare Parameter für eine Praxiseinsatz ausreichend zu evaluieren (Kosch, Hassib, Buschek, & Schmidt, 2018).

Aktuelle Reviews zur objektiven Messung mentaler Beanspruchung mit physiologischen Maßen (z.B. Charles & Nixon, 2019; Cowley et al., 2016) enthalten ferner keine empirischen Nachweise bezüglich der zeitlichen Anwendungseffizienz von Maßen der okulomotorischen Aktivität im Vergleich zu anderen objektiven Methoden, weswegen den hier vorgestellten Maßen keine hohe zeitliche Anwendungseffizienz attestiert werden kann.

3.6.2.3 Hauptaufgabenleistungsmaße

Als Hauptaufgabenleistungsmaße werden alle Maße verstanden, die helfen die Genauigkeit und Korrektheit der Systemnutzung objektiv zu bewerten (Cain, 2007; F. Chen et al., 2016). Dementsprechend können alle Operationalisierungen der Effektivität, einem Leistungskriterium zur Beurteilung intuitiver Benutzung und zentralem pragmatischen Merkmal intuitiver Benutzung (siehe Teilabschnitt 2.2.2), als Hauptaufgabenleistungsmaße bezeichnet werden. Wie in Teilabschnitt 2.1.1 angesprochen, liegt der Effektivität als Leistungskriterium intuitiver Benutzung die Annahme zugrunde, dass mit zunehmender mentaler Beanspruchung diese ab einem gewissen kritischen Punkt abnimmt (siehe F. Chen et al., 2016; Eggemeier et al., 1991; Longo & Leva, 2017; Wu & Li, 2013; Young et al., 2015).

Hauptaufgabenleistungsmaße weisen laut einer Reihe empirischer Arbeiten (z.B. Cain, 2007; F. Chen et al., 2016; Eggemeier et al., 1991; Longo & Leva, 2017) eine ausreichende wissenschaftliche Güte als Maß für mentale Beanspruchung bezüglich der in Teilabschnitt 3.2.5 vorgestellten nicht formalen Kriterien (insbesondere Sensitivität und Diagnostizität) auf. Darüber hinaus konnte bereits für die summative Evaluation intuitiver Benutzung auch ein erster empirischer Nachweis für die Kriteriumsvalidität von Hauptaufgabenleistungsmaßen (d.h. operationalisiert durch Anteil der Fehler und Anteil korrekt gelöster Aufgaben) anhand eines Vergleichs mit dem subjektiven Quasi-Außenkriterium QUESI (d.h. Anteil korrekt gelöster Aufgaben mit QUESI-Gesamt: $r = .56$; Anteil der Fehler mit QUESI-Gesamt: $r = -.46$) bei der Evaluation von Fahrkartenautomaten erbracht werden (Horn, 2008). Ergebnisse bezüglich der Objektivität und Reliabilität fehlen zwar bei Horn (2008), jedoch kann laut Lienert und Raatz (1998) die Prüfung der Validität in gewissem Maße von der Überprüfung der übrigen Gütekriterien entbinden. Für genauere Informationen, wie beispielsweise Korrelationen bezüglich der Subskalen des Quasi-Außenkriteriums QUESI, wird an dieser Stelle auf Horn (2008) verwiesen.

Trotz erster positiver Befunde bezüglich der wissenschaftlichen Güte von Hauptaufgabenleistungsmaßen im Forschungsbereich zu intuitiver Benutzung, sind einige Limitationen bei deren Einsatz bekannt. Intuitive Benutzung lässt sich oftmals nicht alleine über Hauptaufgabenleistungsmaße bewerten, da hierbei kognitive Strategieänderungen bei der Aufteilung der mentalen Ressourcen nicht abgebildet werden können. G. F. Wilson (2005) merkt hierzu an, dass man anhand der reinen Performance nicht automatisch auf den mentalen Zustand des Nutzers schließen kann, da Menschen aufgrund ihrer Adaptionsfähigkeit eine erhöhte mentale Aufgabenbelastung durch eine größere Anstrengung kompensieren können. Insbesondere in Situationen, in denen der Nutzer nur geringer mentaler Belastung ausgesetzt ist und diese dementsprechend durch vermehrte Anstrengung kompensieren kann, ist bei der Interpretation von Effektivität Vorsicht geboten (Cain, 2007; F. Chen et al., 2016; Eggemeier et al., 1991; Hockey, 1997; Wierwille & Eggemeier, 1993).

Umgekehrt, erlauben es solche Adaptionsstrategien (z.B. Unterlassen von Reaktionen oder Warteschlangenbildung) dem Nutzer eine verminderte Performance aufgrund zu hoher mentaler Beanspruchung zu verhindern (Cain, 2007; F. Chen et al., 2016; Eggemeier et al., 1991; Hockey, 1997; Wierwille & Eggemeier, 1993). Dieser Aspekt sollte ebenfalls berücksichtigt werden, wenn Effektivität zur Beurteilung intuitiver Benutzung herangezogen wird. In der Literatur zu mentaler Beanspruchung finden sich hierzu mehrere empirische

Nachweise, die in einem Review von F. Chen et al. (2016) ausführlich diskutiert werden. O'Donnell und Eggemeier (1986) konnten beispielsweise zeigen, dass Hauptaufgabenleistungsmaße Änderungen in der mentalen Beanspruchung nicht abbilden können, wenn diese zu gering ist, da dieser Bereich als „angemessene" Aufgabenleistung gilt und sich deswegen nicht in ganzheitlichen Hauptaufgabenleistungsmaßen, wie Anteil korrekt gelöster Aufgaben, widerspiegelt. Der hohe mentale Belastungsbereich stellt ebenfalls ein Problem für Hauptaufgabenleistungsmaße dar, da dort der Nutzer so überfordert ist, dass es zu einem Erschöpfungszustand kommt und Änderungen in der mentalen Beanspruchung zu keinem Unterschied mehr in der Leistung des Nutzers führen (F. Chen et al., 2016).

S. G. Hart und Wickens (1990) konnten in diesem Zusammenhang zeigen, dass Hauptaufgabenleistungsmaße Änderungen in der mentalen Beanspruchung am besten im mittleren Bereich abbilden können. Ferner verrät das aktuelle Leistungsniveau des Nutzers dem Evaluator nicht, inwiefern eine Anpassungsreaktion aufgrund der mentalen Beanspruchung, verursacht durch die Aufgabe bzw. durch die inhärente Aufgabenschwierigkeit, stattgefunden hat (G. F. Wilson, 2005) oder ob sie generell auf Motivationsprobleme bei der Aufgabenbearbeitung zurückzuführen ist (Gopher & Donchin, 1986). Nur unter Bedingungen, unter denen kein offensichtlicher Leistungsabfall aufgrund zu hoher mentaler Beanspruchung zu verzeichnen ist, belegen subjektive und physiologische Maße, dass sich die mentale Beanspruchung durch die Performance in der Hauptaufgabe abbilden lässt (Cain, 2007).

Dementsprechend sollten Hauptaufgabenleistungsmaße immer nur in Ergänzung und nicht ausschließlich für die objektiv summative Evaluation intuitiver Benutzung eingesetzt werden. Dies wurde bereits bei der Formulierung der Messdefinition für intuitive Benutzung thematisiert. Effektivität wird in der im Rahmen dieser Arbeit abgeleiteten Messdefinition nicht nur als Korrelat mentaler Beanspruchung betrachtet, sondern als zentrales pragmatisches Merkmal intuitiver Benutzung (siehe Teilabschnitt 2.2.2). Im Forschungsbereich zu mentaler Beanspruchung raten hierzu Wickens et al. (2015), dass man sich beim Einsatz von Hauptaufgabenleistungsmaßen immer bewusst sein soll, dass man damit mit hoher Wahrscheinlichkeit nur Schwankungen der mentalen Beanspruchung im mittleren Bereich abbilden kann. Man muss laut Wickens et al. (2015) ferner darauf achten, dass die Leistung des Nutzers nicht durch Stress oder Ermüdungserscheinungen beeinflusst wird, weswegen man die Dauer der untersuchten Aufgaben beim Einsatz von Hauptaufgabenleistungsmaßen dementsprechend kurzhalten sollte. Schließlich sollte man sich darüber im Klaren sein, dass Hauptaufgabenleistungsmaße nichts bringen, wenn man nicht durch andere kontinuierliche objektive Methoden feststellen kann, inwiefern der Nutzer Kompensationsstrategien während der Aufgabenbearbeitung angewendet hat.

Sowohl aktuelle Reviews zur objektiven Messung intuitiver Benutzung mit Hauptaufgabenleistungsmaßen (z.B. Blackler, 2018; Blackler & Hurtienne, 2007; Blackler & Popovic, 2015) als auch die Arbeit von Horn (2008) enthalten außerdem keine empirischen Nachweise bezüglich der zeitlichen Anwendungseffizienz von Hauptaufgabenleistungsmaßen im Vergleich zu anderen objektiven Methoden, weswegen Hauptaufgabenleistungsmaßen keine hohe zeitliche Anwendungseffizienz attestiert werden kann.

3.6.2.4 Zweitaufgabenleistungsmaße

Da durch die eigentliche Systemnutzung (d.h. Hauptaufgabe) nicht alle kognitiven Ressourcen des Nutzers gebunden werden, sind diese für andere Aktivitäten (d.h. Zweitaufgabe) nutzbar (Kerkau, 2006). Mithilfe von Zweitaufgabenleistungsmaßen lässt sich deswegen die übrige mentale Kapazität des Nutzers erschließen, indem er eine Zweitaufgabe parallel zu seiner Hauptaufgabe durchführt (Cain, 2007; F. Chen et al., 2016; Longo, 2014). Anhand der Performance (d.h. Maße der Effektivität oder Effizienz) der Zweitaufgabe können so die Interaktion zwischen Haupt- und Zweitaufgabe und damit die mentale Beanspruchung des Nutzers bei der Hauptaufgabe bewertet werden. Das Zurückgreifen auf Zweitaufgabenleistungsmaße anstatt auf Hauptaufgabenleistungsmaße ist insbesondere dann sinnvoll, wenn man damit rechnet, dass der Nutzer mithilfe von Adaptionprozessen eine erhöhte mentale Belastung so kompensieren kann, dass sich Änderungen in der mentalen Beanspruchung nicht auf Hauptaufgabenleistungsmaße auswirken können (Cain, 2007; F. Chen et al., 2016; Young et al., 2015).

Mithilfe von Zweitaufgabenleistungsmaßen können subtilere Änderungen in der mentalen Beanspruchung als mit Hauptaufgabenleistungsmaßen festgestellt werden, da man mit Zweitaufgaben dafür sorgen kann (d.h. kann die Schwierigkeit der Zweitaufgaben entsprechend anpassen), dass die Leistung der Hauptaufgabe in einem mittleren Bereich stattfindet, in dem man Änderungen auch sensibel genug wahrnehmen kann (Cain, 2007). Laut Cain (2007) muss eine Zweitaufgabe, um mentale Beanspruchung überhaupt abbilden zu können, die folgenden drei Anforderungen erfüllen:

- Die Zweitaufgabe darf sich nicht in die Hauptaufgabe einmischen (d.h. geringe Intrusivität), sollte aber dieselben kognitiven Ressourcen (z.B. visuelle Ressourcen) wie die Hauptaufgabe verwenden, um so viel wie möglich von der gesamten mentalen Beanspruchung abbilden zu können.
- Die Zweitaufgabe muss leicht erlernbar sein.
- Die Zweitaufgabe muss selbstbestimmt sein (d.h. leicht zu unterbrechen oder zu verzögern).

Das *Zweitaufgaben-Paradigma* oder Sekundäraufgabenparadigma lässt sich dabei weiter in das *Paradigma der Nebenaufgabe* (d.h. Subsidiary Task Paradigm, Auxiliary Task Paradigm) und in das *Paradigma der Belastungsaufgabe* (d.h. Loading Task Paradigm) aufteilen. Beide Paradigmen haben jedoch das gleiche Ziel, nämlich die mentale Beanspruchung mithilfe einer Zweitaufgabe messen zu können, ohne dass dies durch Adaptionsstrategien verhindert werden kann (Cain, 2007).

Beim Paradigma der Nebenaufgabe wird die Performance der Hauptaufgabe konstant gehalten, wohingegen beim Paradigma der Belastungsaufgabe die Performance der Zweitaufgabe konstant gehalten wird. Ersteres Paradigma ist dabei am weitesten verbreitet und wird üblicherweise auch mit dem Begriff Zweitaufgaben-Paradigma assoziiert. Nutzer werden hier instruiert eine gleichermaßen gute Performance bei der Erledigung der Hauptaufgabe beizubehalten, unabhängig davon wie schwierig ihre Aufgabe (mit dem System) durch das Hinzukommen der Nebenaufgabe insgesamt wird. Die Veränderung in der Performance der Zweitaufgabe wird dann gemessen und dient als Indikator für die mentale

Beanspruchung des Nutzers unter verschiedenen Beanspruchungsbedingungen bzw. Anforderungen in der Aufgabenkomplexität (Cain, 2007).

Das Paradigma der Belastungsaufgabe sieht vor, dass der Nutzer die Performance der Zweitaufgabe durchgängig gleich halten muss, was vorsätzlich zu einer schlechteren Performance bei der Hauptaufgabe führt. Wie bereits oben beschrieben können Adaptionsstrategien nur angewendet werden, wenn der Nutzer dafür die nötigen mentalen Ressourcen übrighat. Mithilfe des Paradigmas der Belastungsaufgabe kann die Performance der Hauptaufgabe auf ein Niveau gebracht werden, auf dem sie empfindlich auf etwaige Beanspruchungsänderungen bzw. Änderungen in Aufgabenkomplexität reagiert. Der Leistungsabfall der Hauptaufgabe wird gemessen, während die Schwierigkeit der Nebenaufgabe erhöht wird. Im Gegensatz zum Paradigma der Nebenaufgabe, welches die eigentliche Hauptaufgabe nicht beeinflusst und als weniger intrusiv angesehen wird, stellt das Paradigma der Belastungsaufgabe ein intrusives Verfahren dar und kommt deswegen selten zum Einsatz (Cain, 2007).

Egal welches der beiden Paradigmen schlussendlich gewählt wird, man hat bei beiden Paradigmen das Problem, eine für die Hauptaufgabe passende Zweitaufgabe finden zu müssen, die zwar (1) die gleichen kognitiven Ressourcen wie die Hauptaufgabe benötigt, aber (2) nicht mit der Hauptaufgabe interferiert. Nur mit einer möglichst geringen Intrusivität der gewählten Zweitaufgabe kann die mentale Beanspruchung bei der Hauptaufgabe valide und reliabel gemessen werden (Cain, 2007). Um die Passung von Haupt- und Nebenaufgabe zu prüfen, kommen in der HCI verschiedene psychologische Modelle zum Einsatz. Das im Bereich der HCI dafür bekannteste und einflussreichste Modell (Young et al., 2015), ist das *Modell multipler Ressourcen* (Wickens, 1980, 1991, 2002, 2004, 2008; Wickens & Liu, 1988), obwohl dessen Nützlichkeit aus verschiedenen Gründen stark kritisiert wird. Es existieren beispielsweise verschiedene empirische Gegenbelege für zentrale Vorhersagen des Modells (siehe Neumann, 1996) oder es wird die fehlende Berücksichtigung einer taktilen Modalität kritisiert (siehe Boles, Bursk, Phillips, & Perdelwitz, 2007). Unabhängig von dieser Kritik ist das Modell zumindest gut zur Veranschaulichung geeignet, da es aufzeigt, inwiefern sich zwei gleichzeitig auszuführende Aufgaben prinzipiell aufgrund ihres benötigten Kapazitätsniveaus und der damit verbundenen spezifischen Strukturen, Prozesse und Kapazitäten beeinträchtigen können. Für diesen Zweck soll dieses Modell auch im Rahmen dieser Arbeit genutzt und deswegen auf die Einschränkungen nicht weiter eingegangen werden.

Das Modell multipler Ressourcen stellt ein Hybridmodell dar (siehe Abbildung 3.5), welches zum einen Aspekte verschiedener klassischer Kapazitätsmodelle (d.h. Modelle unspezifischer Kapazität), wie das von Kahneman (1973) und das von D. A. Norman und Bobrow (1975) beinhaltet (siehe Goldhammer & Moosbrugger, 2006).

Modelle unspezifischer Kapazität gehen laut Neumann (1996) von einer globalen, unspezifischen und gleichzeitig begrenzten kognitiven Verarbeitungskapazität des Menschen aus, welche auf die verschiedenen Anforderungen der Reizaufnahme, Reizverarbeitung und Handlung durch einen zentralen Prozessor aufgeteilt werden muss (siehe Goldhammer & Moosbrugger, 2006). Genauere Informationen zu Modellen unspezifischer Kapazität sind der Arbeit von Goldhammer und Moosbrugger (2006) zu entnehmen. Zum anderen berücksichtigt das Modell multipler Ressourcen auch Aspekte von klassischen Strukturmodellen (d.h. Modelle spezifischer Kapazität) (z.B. Allport, 1980; Baddeley & Hitch, 1974; Keele,

1973; Kerr, 1973; Navon & Gopher, 1979; Rasmussen, 1983; Welford, 1967), die im Gegensatz zu Kapazitätsmodellen davon ausgehen, dass der Mensch über mehrere voneinander unabhängige, spezialisierte, aber gleichzeitig kapazitätsbegrenzte kognitive Verarbeitungsmechanismen verfügt (Goldhammer & Moosbrugger, 2006; Neumann, 1996). Jedes dieser Module dient laut Wickens (2008) damit der Realisierung einer bestimmten Fertigkeit oder Fähigkeit (z.B. ein Verarbeitungsmechanismus kümmert sich um die kognitive Verarbeitung akustischer und ein anderer um die visueller Informationen).

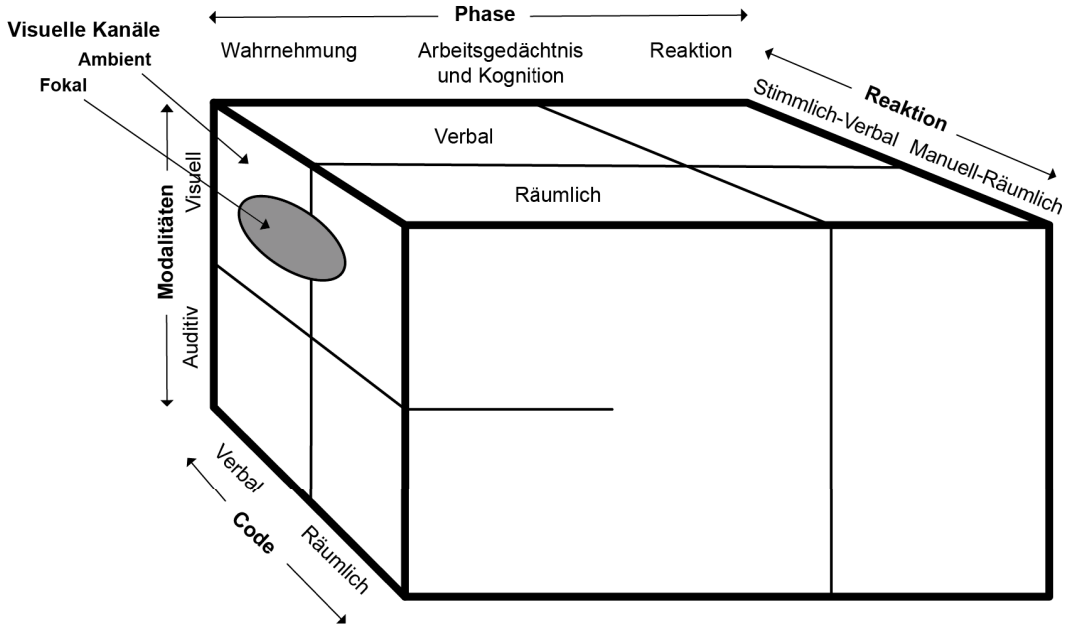


Abbildung 3.5. Modell multipler Ressourcen in Anlehnung an Wickens (2008) unter Berücksichtigung des Arbeitsgedächtnisses von Baddeley (2007) bei der kognitiven Informationsverarbeitung.

Aufgrund der Kapazitätsbegrenzung eines jeden Verarbeitungsmoduls kommt es bei der gleichzeitigen kognitiven Verarbeitung von Aufgaben, wie es bei Evaluation mentaler Beanspruchung anhand von Zweitaufgabenleistungsmaßen der Fall ist, zu einem Wettbewerb um diese Ressourcen und folglich zu wechselseitiger *Interferenz* zwischen der Hauptaufgabe und der Zweitaufgabe (Goldhammer & Moosbrugger, 2006; Neumann, 1996). Das Modell multipler Ressourcen geht zwar auch von einer allgemeinen Limitation des kognitiven Systems aus, postuliert jedoch gleichzeitig, dass, im Gegensatz zu einer einzelnen zentralen Ressource, jeweils spezifische Ressourcen zur Verarbeitung von Informationen benötigt werden, die wiederum über eine beschränkte und eine relativ unabhängige Kapazität verfügen (Wickens, 2008).

Laut Wickens (2008) lassen sich Ressourcen zur Verarbeitung von Informationen anhand von vier kategorischen dichotomen Dimensionen beschreiben (siehe Abbildung 3.5):

- *Verarbeitungsstufen* (d.h. Phase), die in perzeptiv-kognitive Informationsverarbeitung (d.h. Wahrnehmung, Arbeitsgedächtnis und Kognition) und Auswählen bzw.

Ausführen einer Reaktion eingeteilt werden. Laut Wickens (2008) deutet diese Dimension darauf hin, dass wahrnehmende und kognitive Aufgaben verschiedene Ressourcen für die Auswahl und Ausführung zugrunde liegender Aktionen nutzen.

- *Wahrnehmungsmodalitäten* (d.h. Modalitäten), welche in visuelle und auditive Modalitäten aufgeteilt werden. Diese Dimension, welche sich lediglich in der Verarbeitungsstufe der Wahrnehmung befindet, beschreibt, dass die auditive Wahrnehmung andere Ressourcen nutzt, als die visuelle Wahrnehmung (Wickens, 2008).
- *Verarbeitungsweisen* (d.h. Code) können entweder räumlich oder verbal sein. Laut Wickens (2008) beschreibt diese Dimension, dass räumliche Tätigkeiten andere Ressourcen nutzen als verbale Tätigkeiten. Diese Zweiteilung findet während aller Verarbeitungsstufen, nämlich während der Wahrnehmung (siehe Kerr, Condon, & McDonald, 1985), der kognitiven Informationsverarbeitung im Arbeitsgedächtnis (siehe Baddeley, 1992) und beim Ausführen einer Reaktion (d.h. stimmlich verbal vs. manuell räumlich) (siehe Y. Liu & Wickens, 1992), statt.
- *Visuelle Kanäle* unterscheiden zwischen fokalem und ambientem Sehen (Wickens, 2008). Fokales bzw. zentrales Sehen, welches hauptsächlich foveal ist, unterstützt bei der exakten Objektwahrnehmung und ist insbesondere beim Lesen von Text oder Erkennen von Symbolen wichtig, da hierfür eine hohe Sehschärfe benötigt wird. Das ambiente bzw. überwiegend periphere Sehen ist im Gegensatz zum fokalen Sehen für die Wahrnehmung von Orientierung und Bewegung (z.B. Spurhalten auf der Straße beim Autofahren) zuständig (Wickens, 2008). Diese Dimension, welche lediglich in der visuellen Verarbeitung zur Anwendung kommt, postuliert dass für fokales und ambientes Sehen unterschiedliche Ressourcen zum Einsatz kommen (Wickens, 2008).

Je höher die Überlappung der von beiden Aufgaben benötigten Ressourcen bzw. die Interferenz der beiden Aufgaben, umso sensitiver kann die mentale Beanspruchung der Hauptaufgabe durch die Zweitaufgabe widergespiegelt werden (Wickens, 2008; Wickens & Liu, 1988). Unglücklicherweise ist an dieser Stelle auch die Intrusion der Hauptaufgabe durch die Zweitaufgabe am höchsten. Findet umgekehrt keine Interferenz zwischen beiden Aufgaben statt, lässt sich auch keine aus dieser Überlappung resultierende Intrusion feststellen (Wickens, 2008). Das Messen von mentaler Beanspruchung ist jedoch dann auch nicht möglich. Dies stellt ein grundlegendes Problem der Messung von mentaler Beanspruchung mittels Zweitaufgabenleistungsmaßen dar (Damos, 1991). An dieser Stelle sei darauf hingewiesen, dass das Modell multipler Ressourcen zwischen perzeptiv-kognitiver Informationsverarbeitung und der eigentlichen Aktionsausführung trennt. Die mentale Beanspruchung bezieht sich dabei nur auf die perzeptiv-kognitive Informationsverarbeitung (siehe Teilabschnitt 2.2.2), wohingegen sich die physische Arbeitsbelastung auf die eigentlichen Aktionsausführung bezieht. Im Rahmen dieser Arbeit soll sich deswegen auf die Betrachtung der perzeptiv-kognitiven Informationsverarbeitung und dem damit verbundenen Arbeitsgedächtnis beschränkt werden, da sich intuitive Benutzung nur in der mentalen und nicht in der physischen Beanspruchung bei der tatsächlichen Handlungsausführung widerspiegelt (siehe Teilabschnitt 2.1.1).

Wie erwähnt, wird das Modell multipler Ressourcen auch kritisiert, da es eine Vielzahl empirisch schon lange bekannte Interferenzmuster (siehe Ogden, Levine, & Eisner, 1979) nicht erklären kann (Neumann, 1996). Für Doppelaufgaben mit jeweils unterschiedlichen Ausprägungen in den vier kategorialen Dimensionen wird nämlich durch das Modell keinerlei

Intrusion vorhergesagt, was empirisch nicht zu halten ist (Goldhammer & Moosbrugger, 2006; Neumann, 1996). Eine mögliche Erklärung dafür könnte die Tatsache liefern, dass Wickens (1980) in seinem ursprünglichen Modell (d.h. erste Version) keine Aussage trifft, durch welchen Mechanismus die Ressourcen den verschiedenen Aufgaben zugeteilt werden. Bei klassischen Laboruntersuchungen wird dies üblicherweise dadurch unterstützt, dass Nutzern konkret gesagt wird, welche Aufgabe als Haupt- und welche als Nebenaufgabe zu behandeln ist. Es wird auf die Weise direkt eines der beiden vorgestellten Zweiaufgaben-Paradigmen induziert (Tsang, 2006).

Außerhalb des Labors, beispielsweise bei einer summativen Evaluation im Rahmen des menschzentrierten Gestaltungsprozesses, können aber ohne diese expliziten Instruktionen mit dem Modell nicht erklärbare empirische Interferenzmuster auftreten (Neumann, 1996). Wickens (1991) räumt ein, dass sein Modell ja nur Intrusion vorhersagen kann, die aus der Interferenz von Haupt- und Zweitaufgabe resultiert. Es kann beispielsweise auch unabhängig davon Intrusion entstehen, welche aus Einschränkungen bezüglich der menschlichen Eingabe- und Ausgabemechanismen (z.B. Menschen können nicht zwei Positionen gleichzeitig mit ihren Augen betrachten) hervorgeht. Wickens (2008) weist in einer aktuelleren Arbeit darauf hin, dass für die Intrusion auch weitere Phänomene wie ungewollte Ablenkungen wegen Unterbrechungen (z.B. Adler & Benbunan-Fich, 2015; Pankok Jr et al., 2017; Trafton & Monk, 2007), *Cognitive Tunneling* (z.B. Karar & Ghosh, 2014; Thomas & Wickens, 2001; Wickens & Alexander, 2009), die Komplexität der Hauptaufgabe selbst (z.B. Damos, 1991) und *Auditory Preemption* (z.B. Horrey & Wickens, 2004; Wickens & Colcombe, 2007; Wickens, Dixon, & Seppelt, 2005) verantwortlich sind, die nichts mit der Überlappung zu tun haben, sondern vielmehr auf eine nicht optimale Zuteilung von Ressourcen hinweisen.

Laut Lund (2001) ist derartige Kritik nicht überraschend, da das Modell multipler Ressourcen postuliert, ein ganzheitliches Modell zu sein, welches eine vollständige Synthese aus Modellen spezifischer und unspezifischer Kapazität darstellen soll, was impliziert, dass alle jeweiligen Besonderheiten kombiniert werden müssen. Das Modell von Wickens (1980) in seiner ersten Version kombiniert zwar Kapazitätsbeschränkung und multiple Ressourcen miteinander, vergisst aber dabei einen zentralen, in seiner Kapazität beschränkten Prozessor (d.h. zentrale Exekutive) miteinzubeziehen, der für die Allokation von Ressourcen in Form einer Metakognition zuständig ist (siehe Teilabschnitt 2.1.1). Das Modell kann somit keine Aussagen bezüglich der Intrusion treffen, die durch die Allokation der Ressourcen entsteht (siehe Damos, 1991). Eine Reihe von empirischen Arbeiten zeigten jedoch bereits früh (z.B. Baddeley, 1992; A. Miyake & Shah, 1999), dass die Ressourcenverteilung eine Funktion einer zentralen Exekutive ist, weswegen diese zusammen mit dem Konzept des Arbeitsgedächtnisses von Baddeley (1992) mittlerweile im Rahmen einer überarbeiteten zweiten Version des Modells multipler Ressourcen berücksichtigt ist (Wickens, 2008). Das überarbeitete Modell von Wickens (2008) stellt den aktuellen State of the art im HCI-Bereich dar (Matthews et al., 2015; Young et al., 2015) und das Arbeitsgedächtnis ist unter Berücksichtigung von Baddeley (1992) deswegen auch in Abbildung 3.5 bereits berücksichtigt.

Durch die Berücksichtigung einer zentralen Exekutive lässt sich eigentlich kein Unterschied zwischen dem Modell multipler Ressourcen und dem Modell des Arbeitsgedächtnisses bezüglich der zur Messung mentaler Beanspruchung relevanten perzeptiv-kognitiven

Informationsverarbeitung mehr feststellen (Young & Stanton, 2002). Das Arbeitsgedächtnismodell beinhaltet ebenfalls einen zentralen, in seiner Kapazität begrenzten Prozessor (d.h. zentrale Exekutive) mit zwei modalitätsspezifischen, kapazitätsbegrenzten Verarbeitungssystemen (d.h. visuell-räumlicher Notizblock, phonologische Schleife), die nahezu identisch mit den von Wickens (1980, 1991) vorgeschlagenen Verhaltensweisen (d.h. räumlich, verbal) sind. Sie entsprechen somit der von Wickens (1980, 1991) beschriebenen perzeptiven-kognitiven Verarbeitung (Young & Stanton, 2002). Dementsprechend ließ sich das Modell des Arbeitsgedächtnisses auch problemlos in die überarbeitete zweite Version des Modells (siehe Abbildung 3.5) integrieren (Wickens, 2008). Aufgrund der Tatsache, dass die Ressourcenverteilung durch die zentrale Exekutive jedoch noch nicht ausgiebig erforscht ist (siehe Friedman & Miyake, 2017; A. Miyake & Friedman, 2012; A. Miyake et al., 2000; Wongupparaj, Kumari, & Morris, 2015), können auch unter Berücksichtigung einer zentralen Exekutive und des Modells des Arbeitsgedächtnisses innerhalb des Modells multipler Ressourcen die bereits angesprochenen Interferenzmuster (z.B. vorliegende Interferenz obwohl sich Haupt- und Nebenaufgabe bezüglich Verarbeitungsweise, Wahrnehmungsmodalität und Verarbeitungsstufe unterscheiden, siehe Neumann, 1996) nicht vollständig erklärt werden (Wickens, 2008, 2017).

Aufgrund der Tatsache, dass es aktuell kein Modell gibt, das das Ausmaß der Intrusion beim Kombinieren von zwei Aufgaben valide vorhersagen kann (siehe Wickens, 2008, 2017), steht die HCI bei der Evaluation mentaler Beanspruchung mithilfe von Zweitaufgaben ständig vor der Frage, welche Zweitaufgabe mit der gewählten Hauptaufgabe kombiniert werden sollte, um die für die Evaluation mentaler Beanspruchung anhand von Zweitaufgabenleistungsmaßen nötige Interferenz bei möglichst geringer Intrusion zu erzeugen (Cain, 2007). Da im Rahmen dieser Arbeit CUIs im Vordergrund stehen, ist hier insbesondere Literatur zu nennen, die diese Problematik in Zusammenhang mit damit verbundenen Desktop-Systemen schildert (z.B. Bergman, Tene-Rubinstein, & Shalom, 2013; Gwizdka, 2010; Longo & Dondio, 2015). Trotz der Tatsache, dass kein Modell existiert, das die Intrusion „perfekt“ vorhersagen kann, ist man sich in der Forschung sicher, dass, wenn Haupt- und Zweitaufgaben in derselben Modalität dargeboten werden, die Interferenz zwar maximal ist und man deswegen theoretisch den kompletten modalitätsspezifischen Anteil mentaler Beanspruchung erfassen kann, aber dafür auch gleichzeitig die maximale Intrusion in Kauf nehmen muss. Diese Intrusion verhindert letztendlich eine valide, kontinuierliche Messung mentaler Beanspruchung der Hauptaufgabe. Der damit verbundene *Strukturwechseleffekt* wurde sowohl in der HCI (z.B. Goldhammer & Moosbrugger, 2006; Wickens, 1980, 2008, 2017), als auch außerhalb dieses Forschungsfeldes bereits empirisch belegt (z.B. M. A. Cohen, Cavanagh, Chun, & Nakayama, 2012; Ginns, 2005; McIsaac, Lamberg, & Muratori, 2015; McLeod, 1977; Pashler, 1994).

Um trotz dieses Effekts eine geringe Intrusivität zu haben, kombiniert man dementsprechend Aufgaben (siehe Abbildung 3.5), die nicht in derselben Modalität (z.B. Hauptaufgabe visuell, Zweitaufgabe akustisch) liegen, aber entweder räumlich oder verbal verarbeitet werden (Cain, 2007; Damos, 1991; Wickens, 2008). Abhängig von der Hauptaufgabe und deren Ressourcenanforderung kommen dementsprechend eine Vielzahl von verschiedenen Zweitaufgaben zum Einsatz, was eine Standardisierung über mehrere Aufgaben und Domänen enorm erschwert (Cain, 2007). Als Zweitaufgabenleistungsmaße kommen verschiedene Maße der Effizienz (z.B. Reaktionszeiten) und Effektivität (z.B. Genauigkeit, Anzahl der

Fehler) bei unterschiedlichen Tätigkeiten zum Einsatz, wovon unterschiedliche Operationalisierungen in verschiedenen Modalitäten existieren (Cain, 2007).

Es existiert eine hohe Anzahl an Zweitaufgaben in der Literatur, die zum größten Teil in einem aktuellen Review von Gawron (2019) ausführlich (d.h. inklusive Stärken, Schwächen und Anforderungen) diskutiert werden. Dabei lassen sich unter anderem die folgenden Arten von Zweitaufgaben identifizieren:

- Rhythmisches Klopfen: Unter anderem *Tapping Regularity Task* ursprünglich von Wierwille, Rahimi und Casali (1985) und regelmäßiges Klopfen ursprünglich von Y. Miyake, Onishi und Poppel (2004)
- Generierung von Zufallszahlen (z.B. Baddeley, 1966; R. G. Brown & Marsden, 1991; Lysaght, Hill, Dick, Plamondon, & Linton, 1989; Truijens, Trumbo, & Wagenaar, 1976)
- Reaktionszeitmessung bei visuellen Suchaufgaben und Gedächtnisaufgaben: Unter anderem *Sternberg Task* ursprünglich von Sternberg (1969) und *Choice Reaction Time Secondary Task* ursprünglich von Lysaght et al. (1989)
- Verbale Verschleierung (z.B. Lysaght et al., 1989; Savage, Wierwille, & Cordes, 1978; Wielgus & Harvey, 1988)
- Räumliches Denken (z.B. Lysaght et al., 1989; Thurstone, 1944; Vidulich & Tsang, 1985)
- Zeitschätzungen (z.B. S. G. Hart, McPherson, & Loomis, 1978; Wierwille & Connor, 1983; Wierwille et al., 1985)
- Ortungsaufgaben: Unter anderem *Critical Tracking Task* ursprünglich von Jex, McDonnell und Phatak (1966) und *Pursuit Tracking Task* ursprünglich von Noble, Fitts und Warren (1955)
- Card-Sorting (z.B. Courtney & Shou, 1985; Lysaght et al., 1989; Murdock Jr, 1965)
- Mentale Arithmetik (z.B. Lysaght et al., 1989; Ramacci & Rota, 1975)
- Lesen (z.B. Savage et al., 1978; Wierwille & Gutmann, 1978)
- Ablenkungen (z.B. Drory, 1985; Lysaght et al., 1989; Zeitlin, 1995)
- Klassifizierung (z.B. Damos, 1985; Kobus et al., 1986; Lysaght et al., 1989)
- Lexikalische Entscheidungen (z.B. Lysaght et al., 1989)

So wie bei allen Maßen, die entwickelt wurden, um die mentale Beanspruchung erfassen zu können, kamen zur Sicherstellung der wissenschaftlichen Güte von Zweitaufgabenleistungsmaßen als Maß für mentale Beanspruchung überwiegend die in Teilabschnitt 3.2.5 vorgestellten nicht formalen Gütekriterien (insbesondere Sensitivität und Diagnostizität) zum Einsatz, wobei auch einige Zweitaufgaben in ihrer wissenschaftlichen Güte anhand der genannten psychometrischen Hauptgütekriterien *Objektivität*, *Reliabilität* und *Validität* bewertet wurden. Für diese Informationen sei an dieser Stelle auf das ausführliche Review von Gawron (2019) und die darin referenzierten Arbeiten verwiesen, da eine Diskussion den Rahmen dieser Arbeit sprengen würde. Obwohl die wissenschaftliche Güte einiger Zweitaufgaben außer Frage steht, gefährdet ein größeres methodologisches Problem die generelle wissenschaftliche Güte des Zweitaufgabenparadigmas. Das Problem bei

all diesen Zweitaufgaben ist, dass man damit immer nur spezifische Aspekte der mentalen Gesamtbeanspruchung des Nutzers messen kann (Park & Brünken, 2015). Wählt man beispielsweise für die Evaluation einer visuellen, räumlichen Hauptaufgabe eine akustische, räumliche Zweitaufgabe, misst man damit auch nur den räumlichen Anteil innerhalb der mentalen Gesamtbeanspruchung. Beinhaltet die Hauptaufgabe zusätzlich noch verbal-sprachliche Aspekte, müssen diese auch in gleichem Maße durch die Zweitaufgabe berücksichtigt werden, was das Unterfangen extrem anspruchsvoll macht. Wie bereits angesprochen, fehlt hier ein passendes Modell, mit dem der Anwender eine möglichst geringe intrusive, aber gleichzeitig nötige Interferenz erzeugende Zweitaufgabe auswählen kann. Cain (2007) warnt deswegen die Kombination von Haupt- und Zweitaufgabe nur „will-he, nil-he“ (d.h. willkürlich) durchzuführen, ist sich dabei aber der immensen Schwierigkeit bei der Auswahl der richtigen Zweitaufgabe bewusst.

Zur Veranschaulichung dieser Schwierigkeit soll an dieser Stelle eine Studie von Gwizdka (2010) exemplarisch herangezogen werden, der einen Stroop-Test als Zweitaufgabe für die summative Evaluation mentaler Beanspruchung vorschlägt. Er nutzte diesen Stroop-Test zur Messung der mentalen Beanspruchung bei zwei Websuchmaschinen (d.h. Google vs. Alvis). Bei diesem Stroop-Test wurden den Probanden visuell Farbwörter in einem Pop-up-Fenster am Bildschirmrand dargeboten. Passte die Farbe nicht zum Farbwort, mussten die Probanden mit der linken Maustaste auf das Pop-up-Fenster klicken. Bei einer Übereinstimmung von Farbe und Farbwort kennzeichneten die Probanden dies mit einem Klick auf die rechte Maustaste. Im Anschluss wurde die Reaktionszeit als Zweitaufgabenleistungsmaß ausgewertet. Beurteilt man den von Gwizdka (2010) eingesetzten Stroop-Test hinsichtlich der Modalitätsdimension, ist er diesbezüglich auf jeden Fall als geeignete Zweitaufgabe im vorliegenden Szenario zu bewerten, da er die nötige Interferenz erzeugt. Er ist somit theoretisch in der Lage den kompletten modalitätsspezifischen Anteil der mentalen Beanspruchung abzubilden, der durch die visuelle Verarbeitung verursacht wird. Er erfasst dadurch auch jegliche Schwankungen im Ressourcenverbrauch, die bei räumlichen und verbalen Aktivitäten im Arbeitsgedächtnis auftreten (siehe Abbildung 3.5). Obwohl der von Gwizdka (2010) genutzte Stroop-Test somit theoretisch als Zweitaufgabe für die Evaluation der Systemnutzung geeignet ist, besitzt er Einschränkungen.

Gwizdka (2010) bot den Stroop-Test in zufälligen Zeitintervallen (d.h. alle 15 bis 29 Sekunden) für einen zufälligen Zeitraum (d.h. jeweils 5 bis 9 Sekunden) dar. Diese Art der Darbietung machte für Gwizdka (2010) Sinn, da er sich der Intrusion seiner Methode bewusst war und er diese dadurch möglichst gering halten wollte. Aufgrund der Darbietung beider Aufgaben in derselben visuellen Modalität ist die Intrusion generell hoch, was durch die intervallbasierte Darbietung etwas reduziert wurde. Eine vollständig parallele Darbietung würde zwar eine kontinuierliche Aussage bezüglich der mentalen Beanspruchung des Nutzers in nahezu Echtzeit ermöglichen, aber auch zu vielen unerwünschten bewussten Unterbrechungen der Hauptaufgabe und hoher Intrusion führen (z.B. Drugge & Witt, 2006; McFarlane & Latorella, 2002; Nilsson, Drugge, Liljedahl, Synnes, & Parnes, 2005). Die Intrusion ist durch die zufälligen Unterbrechungen bei Gwizdka (2010) jedoch immer noch vorhanden und nicht eliminiert. Da beim Stroop-Test von Gwizdka (2010) die automatische Reaktion sofort das Farbwort zu benennen, ohne dessen eigentliche Farbe zu berücksichtigen, durch kognitives Abkoppeln unterdrückt werden muss (d.h. Unterdrücken der unbewussten Reaktion von Typ 1 Prozessen), handelt es sich bei seiner Zweitaufgabe wie bei jedem Stroop-Test um eine Inhibitionsaufgabe (siehe Stroop, 1935). Wie bereits in

Kapitel 2 erläutert, gelten Inhibitionsprozesse als gute Indikatoren für Prozesse der zentralen Exekutive (J. D. Cohen et al., 1997; A. Miyake et al., 2000; Stanovich et al., 2014) und können demnach als universelle Indikatoren der gesamten mentalen Beanspruchung betrachtet werden (siehe Teilabschnitt 2.2.2).

Der von Gwizdka (2010) genutzte Stroop-Test kann diesen Vorteil aufgrund seiner Modalitätsabhängigkeit und der damit verbundenen hohen Intrusion aber nicht vollständig ausschöpfen. Um das Potential vollständig nutzen zu können, hätte Gwizdka (2010) seinen Stroop-Test beispielsweise akustisch darbieten müssen (d.h. akustisch dargebotene Wörter, wie hoch/niedrig, laut/leise oder schnell/langsam sind entweder kongruent oder nicht kongruent mit der Tonhöhe, der Lautstärke oder der Dauer ihrer Darbietung) (siehe G. Cohen & Martin, 1975; Shor, 1975). Obwohl die Intrusivität bei einem akustischen Stroop im Szenario von Gwizdka (2010) wahrscheinlich zu vernachlässigen ist, besitzt ein akustischer Stroop wiederum Einschränkungen, die seine wissenschaftliche Güte beeinträchtigen können. Beispielsweise kann die Performance hier stark von der Musikalität der Probanden abhängen (z.B. Y. Miyake et al., 2004), der allgemeinen Hörfähigkeit (z.B. Jerger et al., 1993), vom Geschlecht des Sprechers (z.B. E. J. Green & Barber, 1981), der Tonqualität (z.B. E. J. Green & Barber, 1983), der Reiz-Reaktions-Kompatibilität (z.B. McClain, 1983) und sogar von der Händigkeit (d.h. welche Hand wird bevorzugt für Tätigkeiten verwendet) der Probanden (z.B. Morgan & Brandt, 1989).

Mithilfe des oben beschriebenen Beispiels sollte auf einfache Art und Weise veranschaulicht werden, was man alles bei der Auswahl einer Zweitaufgabe bedenken muss. Obwohl akustische Stroop-Aufgaben ähnlich wie visuelle Stroop-Aufgaben Inhibitionsprozesse abbilden können (MacLeod, 1991), verhindern verschiedene Einschränkungen ihren Einsatz als verhältnismäßig universelle Zweitaufgabe für die Evaluation mentaler Beanspruchung. Darüber hinaus kann das aktuelle Review von Gawron (2019) keine empirischen Nachweise bezüglich der zeitlichen Anwendungseffizienz von aktuellen Zweitaufgabenleistungsmaßen im Vergleich zu anderen objektiven Methoden liefern, weswegen aktuellen Zweitaufgabenleistungsmaßen keine hohe zeitliche Anwendungseffizienz attestiert werden kann. Nach meinem Wissen existiert im HCI-Bereich noch keine universelle Zweitaufgabe, der es gleichermaßen gelingt, die nötige Interferenz zu erzeugen und dabei möglichst wenig intrusiv zu sein. Die im Rahmen dieser Arbeit entwickelte Zweitaufgaben *IntuiBeat-S* soll diese Lücke schließen und so für die summative Evaluation intuitiver Benutzung mit hoher zeitlicher Anwendungseffizienz im Anwenderprojekt eingesetzt werden können.

3.6.2.5 Limitationen von objektiven Methoden

Wie zu Beginn dieses Abschnitts beschrieben, kann intuitive Benutzung anhand der mentalen Beanspruchung objektiv bewertet werden. Tabelle 3.3 fasst die Ergebnisse der eben vorgestellten objektiven Methoden bezüglich der wissenschaftlichen Güte (d.h. anhand der wissenschaftlichen formalen Gütekriterien *Objektivität*, *Validität* und *Reliabilität*) und der zeitlichen Anwendungseffizienz zusammen. An dieser Stelle sei angemerkt, dass innerhalb der Tabelle ausschließlich empirische Nachweise der Validität zusammengefasst werden, die direkt im Forschungsfeld zu intuitiver Benutzung erfolgten und nicht lediglich im Forschungsbereich zu mentaler Beanspruchung erbracht wurden. Bezüglich Objektivität und Reliabilität werden in der Tabelle 3.3 auch empirische Nachweise zusammengefasst, die

aus der allgemeinen Forschung zur mentalen Beanspruchung stammen, wenn die objektive Methode aus diesem Feld für die summative Evaluation intuitiver Benutzung entsprechend übernommen wurde. Des Weiteren konnte allen identifizierten objektiven Methoden in empirischer Hinsicht keine hohe zeitliche Anwendungseffizienz im Vergleich zu anderen objektiven Maßen attestiert werden, weil in der Literatur entsprechende empirische Belege dafür fehlen.

Als Verhaltensmaße kommen im Forschungsbereich zu intuitiver Benutzung die Q-Methode von Asikhia (2015) und die CHAI-Methode von Reinhardt et al. (2018) zur objektiven summativen Evaluation intuitiver Benutzung zum Einsatz. Die wissenschaftliche Güte konnte von beiden Methoden bestätigt werden, wobei im Vergleich zur Q-Methode die Konstruktvalidität der CHAI-Methode bereits im CUI-Bereich sichergestellt werden konnte. Bei der Q-Methode wurde lediglich die Kriteriumsvalidität bei der Evaluation von einfachen Weckern gezeigt (siehe Tabelle 3.3). Nichtsdestotrotz können beide Maße aufgrund der Tatsache, dass bereits empirische Nachweise bezüglich der formalen wissenschaftlichen Gütekriterien vorliegen, im Rahmen des Projekts 3D-GUIde zur summativen Evaluation intuitiver Benutzung eingesetzt werden. Empirische Nachweise bezüglich der zeitlichen Anwendungseffizienz liegen jedoch für beide Methoden auf Basis aktueller Reviews (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) und der in den entsprechenden Teilabschnitten genannten Arbeiten nicht vor.

Des Weiteren kann das Ausmaß an intuitiver Benutzung auch über die mentale Beanspruchung objektiv anhand von verschiedenen physiologischen Maßen quantifiziert werden. Innerhalb dieses Teilabschnitts wurden zu diesem Zweck Maße der Herz- und Kreislaufaktivität, der Gehirnaktivität, der elektrodermalen Aktivität, der respiratorischen Aktivität und der okulomotorischen Aktivität vorgestellt. Wie, die in diesem Abschnitt vorgestellten, physiologischen Maße aus den aktuellen Reviews von Cowley et al. (2016) und Charles und Nixon (2019) zeigen sollten, kommen aufgrund des technischen Fortschritts immer mehr physiologische Maße in Labor und Feld zum Einsatz. Obwohl die wissenschaftliche Güte verschiedener physiologischer Maße bezüglich der in Teilabschnitt 3.2.5 genannten, nicht formellen Gütekriterien (insbesondere Sensitivität und Diagnostizität) bestätigt werden konnte, fehlt die Sicherstellung der wissenschaftlichen Güte bezüglich etablierter, psychometrischer, formaler, wissenschaftlicher Hauptgütekriterien (siehe Abschnitt 3.6) (siehe Cain, 2007; Charles & Nixon, 2019; F. Chen et al., 2016; Cowley et al., 2016).

Darüber hinaus schlussfolgerten Cowley et al. (2016), dass die aufgrund des technischen Wandels hinzukommenden Messmöglichkeiten auch ihre Schattenseiten mit sich bringen. Bei der Messung von physiologischen Daten wird durch das Messinstrument selbst immer etwas Umgebungsrauschen (d.h. Noise) erzeugt, welches bei neuen Verfahren (z.B. Messung der Vasokonstriktion aus der Ferne per Webcam, siehe Bousefsaf et al., 2014) noch höher liegt (d.h. schlechteres Signal-Rausch-Verhältnis). Die technische Umsetzung der in diesem Teilabschnitt vorgestellten Methoden wurde zwar nicht besprochen (siehe Cowley et al., 2016), augenscheinlich sollte aber dennoch klar sein, dass hier die wissenschaftliche Güte sehr stark mit dem zeitlichen Aufwand zusammenhängt, welcher für die Durchführung und Analyse der Daten, sowie die Auswahl der Hardware und der Algorithmen aufgewendet wird (Cowley et al., 2016). Cowley et al. (2016) macht für den hohen zeitlichen Aufwand von physiologischen Maßen hier insbesondere die Tatsache verantwortlich, dass

physiologische Reaktionen des Körpers nicht ausschließlich von einer einzigen Quelle verursacht werden (z.B. erhöhte Herzfrequenz bildet nicht notwendigerweise nur die erhöhte mentale Beanspruchung durch die Systemnutzung ab). Jede physiologische Reaktion korreliert stattdessen mit verschiedenen psychophysischen Zuständen und Phänomenen. Dieser Aspekt wirkt sich, neben der verwendeten Messmethode, selbst ebenfalls schlecht auf das Signal-Rausch-Verhältnis von physiologischen Maßen aus, da dadurch mehrere Signale zur selben Zeit und auf derselben Frequenz auftauchen können bzw. das Signal von anderen Faktoren (z.B. emotionaler Zustand) gestört wird (Cowley et al., 2016).

Tabelle 3.3. *Wissenschaftliche Güte bezüglich der formalen Gütekriterien Objektivität, Reliabilität und Validität, sowie bezüglich der zeitlichen Anwendungseffizienz objektiver Evaluationsmethoden für intuitive Benutzung (anhand von Hauptaufgabenleistungsmaßen und mentaler Effizienz).*

Maß	Objektivität	Reliabilität	Validität	Zeitliche Anwendungs-effizienz
Q	Beurteiler-übereinstimmung (Asikhia, 2015)	✗	Innere Kriteriumsvalidität (Asikhia, 2015)	✗
CHAI	Beurteiler-übereinstimmung (Reinhardt, Kuge, & Hurtienne, 2018)	✗	Konstruktvalidität anhand konvergenter und divergenter Validität (Horn, 2008; Reinhardt, Kuge, & Hurtienne, 2018)	✗
Physio. Maße	✗	✗	✗	✗
Hauptaufgaben	✗	✗	Innere Kriteriumsvalidität (Horn, 2008)	✗
Zweitaufgaben	Bestimmte Zweitaufgaben (Gawron, 2019)	Bestimmte Zweitaufgaben (Gawron, 2019)	✗	✗

Wie schon alleine an der Tatsache erkannt werden kann, dass sich in diesem Teilabschnitt bei der Vorstellung der einzelnen physiologischen Maße eine Vielzahl von referenzierten Studien wiederholen, hat es sich wegen der schlechten Zuordenbarkeit von Reaktion und Ursache etabliert, mentale Beanspruchung immer mit mehreren physiologischen Maßen gleichzeitig zu messen (Charles & Nixon, 2019; F. Chen et al., 2016; Cowley et al., 2016). Charles und Nixon (2019) resümieren in diesem Zusammenhang, dass aktuell kein physiologisches Maß existiert, das man ohne die Kombination mit anderen physiologischen Maßen für die summative Evaluation mentaler Beanspruchung einsetzen sollte. Aufgrund des fehlenden formalen Gütenachweises bei physiologischen Maßen bezüglich der formalen Gütekriterien *Objektivität*, *Reliabilität* und *Validität* lässt sich auch kein physiologisches Maß zur objektiven summativen Evaluation intuitiver Benutzung im Rahmen des Pro-

jekts 3D-GUIde einsetzen. Empirische Nachweise bezüglich der zeitlichen Anwendungseffizienz liegen für physiologische Maße im Vergleich zu anderen objektiven Maßen auf Basis aktueller Reviews (siehe Charles & Nixon, 2019; Cowley et al., 2016) und der in den entsprechenden Teilabschnitten genannten Arbeiten außerdem nicht vor.

Des Weiteren kann intuitive Benutzung auch mithilfe von Hauptaufgabenleistungsmaßen bewertet werden. Wie bereits in Kapitel 2 dargestellt, ist Effektivität als Leistungskriterium für intuitive Benutzung zu berücksichtigen, da sie jeder Handlung aufgrund deren inhärenten Zielgerichtetheit zugrunde liegt und dementsprechend als zentrales Merkmal intuitiver Benutzung in der im Rahmen dieser Arbeit vorgestellten Arbeitsdefinition berücksichtigt wurde (siehe Teilabschnitt 2.2.2). Die wissenschaftliche Güte von Maßen der Effektivität bzw. Hauptaufgabenleistungsmaßen konnte im Forschungsbereich zu intuitiver Benutzung bereits in Form einer Kriteriumsvalidierung empirisch nachgewiesen werden (siehe Tabelle 3.3). Nichtsdestotrotz wurde in diesem Zusammenhang auch erläutert, dass Hauptaufgabenleistungsmaße immer nur in Kombination mit anderen Maßen zur Evaluation intuitiver Benutzung genutzt werden sollen. Ohne zusätzliche Maße können mit Hauptaufgabenleistungsmaßen keine Rückschlüsse auf die mentale Beanspruchung während der eigentlichen Nutzung getroffen werden, was zu einer unzureichenden Erfassung des Ausmaßes intuitiver Benutzung führen kann. Die zusätzliche Verwendung weiterer Maße für eine kontinuierliche Erfassung erhöht die Kosten der Methode und senkt auch deren zeitliche Anwendungseffizienz, da dadurch insgesamt mehr Zeit für die Messung aufgewendet werden muss. Aufgrund dieser Tatsache können Hauptaufgabenleistungsmaße zwar zur summativen Evaluation intuitiver Benutzung im Rahmen des Projekts 3D-GUIde eingesetzt werden, dies sollte aber immer in Kombination mit anderen Maßen geschehen, die die intuitive Benutzung des Nutzers auch kontinuierlich während der eigentlichen Aufgabenbearbeitung abbilden können. Empirische Nachweise bezüglich der zeitlichen Anwendungseffizienz liegen für Hauptaufgabenleistungsmaße auf Basis aktueller Reviews (z.B. Blackler, 2018; Blackler & Hurtienne, 2007; Blackler & Popovic, 2015) und den im entsprechenden Teilabschnitt genannten Arbeiten außerdem nicht vor.

Schließlich wurden in diesem Teilabschnitt auch Zweitaufgabenleistungsmaße als letzte Möglichkeit vorgestellt, um intuitive Benutzung anhand der damit verbundenen mentalen Beanspruchung objektiv zu messen. In diesem Teilabschnitt wurde nicht, wie bei den anderen objektiven Methoden, die wissenschaftliche Güte referenzierter Zweitaufgaben im Detail diskutiert und stattdessen auf ein Review von Gawron (2019) verwiesen, sondern das große methodische Problem beim Einsatz von Zweitaufgaben in den Fokus des Teilabschnitts gerückt und dieses anhand eines Fallbeispiels veranschaulicht. Aktuelle Zweitaufgaben haben das Problem, dass es keine universelle Zweitaufgabe gibt, die theoretisch für die Evaluation einer jeden typischen Hauptaufgabe im HCI-Bereich eingesetzt werden kann. Auf Basis des im Teilabschnitt vorgestellten Modells multipler Ressourcen von Wickens (2008) muss sich die Zweitaufgabe mit der Hauptaufgabe die gleichen Ressourcen teilen, um überhaupt mentale Beanspruchung messen zu können. Je stärker die Überschneidung, umso besser bildet die Zweitaufgabe die gesamte mentale Beanspruchung ab und nicht nur Teilaspekte (z.B. modalitätsspezifische mentale Beanspruchung aufgrund visueller Belastung). Die Krux hierbei ist es, dass die Zweitaufgabe dadurch im gleichen Zug eine hohe Intrusion erzeugt und den Nutzer bei der Hauptaufgabe beeinflussen (z.B. ablenken, unterbrechen) kann, was die wissenschaftliche Güte der Methode beeinflusst (Wickens, 2008). Der Teilabschnitt endete mit dem Resümee, dass im HCI-Bereich aktuell

keine Zweitaufgabe existiert, die dieses Problem lösen kann. Aus diesem Grund können aktuelle Zweitaufgabenleistungsmaße nicht zur summativen Evaluation intuitiver Benutzung im Rahmen des Projekts 3D-GUIde genutzt werden. Empirische Nachweise bezüglich der zeitlichen Anwendungseffizienz liegen für Zweitaufgabenleistungsmaße auf Basis eines aktuellen Reviews (siehe Gawron, 2019) und den im entsprechenden Teilabschnitt genannten Arbeiten nicht vor.

Zusammenfassend kann festgehalten werden, dass lediglich die CHAI-Methode die nötige wissenschaftliche Güte besitzt, um als Benchmark für die summativen Evaluation im Rahmen des Projekts 3D-GUIde angesehen zu werden. Da der Nachweis der Kontruktvalidität die wissenschaftliche Güte noch differenzierter sicherstellen kann als der Nachweis der Kriteriumsvalidität, ist sie in diesem Punkt der Q-Methode überlegen (siehe Tabelle 3.3). Da die wissenschaftliche Güte von physiologischen Maßen nicht formal bestätigt werden kann und Hauptaufgabenleistungsmaße nicht alleine für die summative Evaluation intuitiver Benutzung eingesetzt werden sollen, kann die Einstufung der CHAI-Methode als Benchmark durchaus gerechtfertigt werden. Hauptaufgabenleistungsmaße besitzen im Gegensatz zu physiologischen Maßen jedoch eine nachgewiesene wissenschaftliche Güte für die Evaluation intuitiver Benutzung bezüglich formaler wissenschaftlicher Gütekriterien (siehe Tabelle 3.3). Sie können daher für eine Meta-Evaluation einer summativen Evaluationsmethode für intuitive Benutzung als Quasi-Außenkriterium fungieren.

Obwohl aktuelle Zweitaufgaben aufgrund des angesprochenen methodischen Problems nicht zur summativen Evaluation im Rahmen des Projekts 3D-GUIde eingesetzt werden können, bieten Zweitaufgabenleistungsmaße im Gegensatz zu Verhaltensmaßen, physiologischen Maßen und Hauptaufgabenleistungsmaßen jedoch ein großes Potential für die objektiv summative Evaluation intuitiver Benutzung im Rahmen des Projekts 3D-GUIde, da sie

- für die Erfassung mentaler Beanspruchung mit hoher Wahrscheinlichkeit zeitlich anwendungseffizienter als Verhaltensmaße sind, die für die Quantifizierung mentaler Beanspruchung zeitlich aufwendige Analysen (z.B. CHAI-Methode: Videoanalyse mithilfe eines Bewertungsschemas) benötigen.
- für die Erfassung mentaler Beanspruchung lediglich einfache Leistungsmaße (z.B. Reaktionszeit) der Zweitaufgaben analysiert und dafür keine komplexen Algorithmen angewendet werden müssen, um die eigentliche Metrik vom Rauschen trennen zu können, wie dies bei physiologischen Maßen erforderlich ist.
- die Erfassung mentaler Beanspruchung kontinuierlich vornehmen können und nicht nur auf eine retrospektive Bewertung beschränkt sind, wie dies bei Hauptaufgabenleistungsmaßen und subjektiven Methoden der Fall sein kann.

Aufgrund der Tatsache, dass, wie im letzten Teilabschnitt angesprochen, Inhibitionsaufgaben als universelle Zweitaufgaben verwendet werden können, sofern man es schafft, das Problem mit der Modalitätsabhängigkeit zu lösen, kann die im nächsten Kapitel auf Inhibition basierende Zweitaufgabe *IntuiBeat-S* einen neuen Benchmark für die summative Evaluation intuitiver Benutzung darstellen. Dieser Benchmark kann unter Berücksichtigung der Anforderungen des Projekts 3D-GUIde, im Vergleich zum aktuellen Benchmark, der CHAI-Methode, mit hoher zeitlicher Effizienz angewendet werden. Hauptaufgabenleistungsmaße können, bei einer in diesem Zusammenhang erforderlichen Meta-Evaluation

von IntuiBeat-S, als Quasi-Außenkriterium zusammen mit der als Benchmark nutzbaren CHAI-Methode ein echtes Außenkriterium approximieren. Des Weiteren kann eine auf Inhibition basierende Zweitaufgabe *IntuiBeat-F* auch als neuer Benchmark für die formative Evaluation intuitiver Benutzung fungieren, da diese, im Gegensatz zum aktuellen Benchmark, dem Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der Identifikation kritischer Ereignisse unterstützen kann, die mit einer erhöhten mentalen Beanspruchung und damit einem geringeren Ausmaß an intuitiver Benutzung einhergehen. Durch diese gezielte Unterstützung kann IntuiBeat-F auch unter Berücksichtigung der Anforderungen des Projekts 3D-GUIde im Vergleich zum aktuellen Benchmark, dem Nutzertest mit retrospektivem Think-Aloud-Protokoll, mit hoher zeitlicher Effizienz angewendet werden.

3.7 Zusammenfassung

In diesem dritten Kapitel wurde beschrieben, wie sich intuitive Benutzung auf Basis, der im zweiten Kapitel vorgestellten Arbeitsdefinition evaluieren lässt. Zu diesem Zweck wurde zu Beginn des Kapitels aufgezeigt, wo sich die Evaluation intuitiver Benutzung in einen menschzentrierten Gestaltungsprozess verorten lässt. Im Zuge dessen wurde eine Arbeitsdefinition für den Begriff Evaluation und ein für deren Durchführung erforderlicher formaler vierstufiger Prozess beschrieben, der bei der Evaluation eines Evaluationsgegenstandes (z.B. CUI) den Vergleich bestimmter Evaluationskriterien mit Außenkriterien vorsieht. Des Weiteren wurde erörtert, dass mit einer Evaluation verschiedene Evaluationsziele (d.h. analysierend, vergleichend, bewertend) verfolgt werden können und auf Basis der möglichen Evaluationsziele grundsätzlich zwischen einer formativen (d.h. analysierend) und einer summativen (d.h. vergleichend, bewertend) Evaluation unterschieden werden kann.

Letztere erlaubt dabei auch die Evaluation einer Evaluationsmethode selbst, weswegen im Zuge dessen auch der Begriff der Meta-Evaluation eingeführt und die mangelnde formale Meta-Evaluation innerhalb der HCI (damit auch im Forschungsbereich zu intuitiver Benutzung) diskutiert wurde. Hierbei wurde speziell kritisiert, dass Evaluationsmethoden nicht anhand eines Vergleichs mit einem echten Außenkriterium (d.h. weist semantische und theoretische Ähnlichkeit zur untersuchenden Evaluationsmethode auf und ist zusätzlich von höherem Status als diese Evaluationsmethode), sondern lediglich anhand nicht formaler Gütekriterien (z.B. Sensitivität, Diagnostizität) ohne direkten Vergleich mit einem echten Außenkriterium evaluiert werden. Als Gütekriterien werden bei einer Meta-Evaluation die Evaluationskriterien bezeichnet. Da sich echte Außenkriterien für eine Meta-Evaluation oftmals schwer festlegen lassen, wurde der Begriff des Quasi-Außenkriteriums (d.h. weist semantische und theoretische Ähnlichkeit zur untersuchenden Evaluationsmethode auf, ist jedoch nicht von höherem Status als diese Evaluationsmethode) eingeführt, mit dem sich ein echtes Außenkriterium approximieren lässt, wenn davon mehrere zum Einsatz kommen. An dieser Stelle wurde außerdem der Begriff des Benchmarks eingeführt, mit dem ein Quasi-Außenkriterium bezeichnet wird, welches von allen verfügbaren Quasi-Außenkriterien die höchste formal nachgewiesene wissenschaftliche Güte aufweist. Es eignet sich somit am besten für die Evaluation und stellt damit den Benchmark im Forschungsfeld zu intuitiver Benutzung dar.

Im Anschluss wurden die formalen wissenschaftlichen Gütekriterien (d.h. Hauptgütekriterien) für eine formative Evaluationsmethode (d.h. Gründlichkeit, Gültigkeit und Zuver-

lässigkeit) vorgestellt. Aufgrund der Tatsache, dass mit diesen Hauptgütekriterien nur die wissenschaftliche Güte, also die Fähigkeit der Methode wissenschaftlich tragfähige Ergebnisse zu produzieren, abgesichert werden kann und nicht das Anwenderinteresse, also die praktische Güte, wurden zusätzlich noch eine Reihe von formativen Nebengütekriterien vorgestellt. Ein Nebengütekriterium berücksichtigte hierbei auch die aus der Perspektive des Projekts 3D-GUIde geforderte, zeitliche Anwendungseffizienz einer formativen Evaluationsmethode für intuitive Benutzung. Anschließend wurden die formalen wissenschaftlichen Gütekriterien für eine summative Evaluationsmethode (d.h. Objektivität, Reliabilität und Validität) vorgestellt. Da, wie bei den formativen Hauptgütekriterien, damit auch nur die wissenschaftliche Güte der Methode abgesichert werden kann und nicht das Anwenderinteresse, wurden auch hier noch zusätzlich eine Reihe von summativen Nebengütekriterien vorgestellt, wobei ein Nebengütekriterium auch, die aus der Perspektive des Projekts 3D-GUIde geforderte, zeitliche Anwendungseffizienz einer summativen Evaluationsmethode für intuitive Benutzung berücksichtigte.

Daraufhin wurde der aktuelle Forschungsstand bezüglich der formativen Evaluation intuitiver Benutzung unter Berücksichtigung der vorgestellten formativen Hauptgütekriterien und der zeitlichen Anwendungseffizienz bewertet. Hierbei wurde festgestellt, dass ein Nutzertest mit retrospektivem Think-Aloud-Protokoll im Bereich der formativen Evaluation intuitiver Benutzung als Benchmark fungiert und gleichzeitig auch das einzige Quasi-Außenkriterium bei einer zukünftigen Meta-Evaluation zur Sicherstellung der wissenschaftlichen Güte einer formativen Evaluationsmethode darstellt. Im Anschluss wurde auch der aktuelle Forschungsstand bezüglich der summativen Evaluation intuitiver Benutzung unter Berücksichtigung der summativen Hauptgütekriterien und der zeitlichen Anwendungseffizienz bewertet. Hierbei wurden zunächst die im Forschungsfeld zu intuitiver Benutzung verfügbaren subjektiven Methoden diskutiert (d.h. TFQ, QUESI, INTUI, SMEQ/SEA und NASA-TLX) und die SEA-Skala, der QUESI und der NASA-TLX als subjektive Quasi-Außenkriterien für eine zukünftige Meta-Evaluation zur Sicherstellung der wissenschaftlichen Güte einer summativen Evaluationsmethode gewählt. Im Anschluss erfolgte die Diskussion objektiver, im Forschungsfeld zu intuitiver Benutzung verfügbaren Methoden (d.h. Verhaltensmaße, physiologische Maße, Hauptaufgabenleistungsmaße, Zweitaufgabenleistungsmaße), wobei Hauptaufgabenleistungsmaße in Form von Maßen der Effektivität als objektives Quasi-Außenkriterium für eine zukünftige Meta-Evaluation zur Sicherstellung der wissenschaftlichen Güte einer summativen Methode identifiziert werden konnten. Ferner konnte gezeigt werden, dass sich die CHAI-Methode auf Basis der im Feld vorliegenden empirischen Befunde mit dem Nachweis der Konstruktvalidität als Benchmark für die summative Evaluation im Rahmen des Projekts 3D-GUIde qualifiziert und damit auch als solcher bei einer zukünftigen Meta-Evaluation einer summativen Evaluationsmethode fungieren kann.

Aufgrund der Tatsache, dass die CHAI-Methode eine ausgiebige Analyse von Videomaterial für die Anwendung des CHAI-Beurteilungsschemas erfordert und ihre zeitliche Anwendungseffizienz damit nicht als hoch einzustufen ist, wurde abschließend anhand des Modells multipler Ressourcen und eines Fallbeispiels argumentiert, dass Zweitaufgaben zwar ein hohes Potential für die Evaluation intuitiver Benutzung besitzen, es jedoch aktuell keine universelle Zweitaufgabe gibt, die zur summativen Evaluation von 3D-CUI-Interaktionslösungen im Projekt 3D-GUIde einsetzt werden kann. Hier wurde begründet, dass eine auf Inhibition basierende Zweitaufgabe einen neuen Benchmark für die for-

mative und summative Evaluation intuitiver Benutzung darstellen kann, der mit hoher zeitlicher Anwendungseffizienz im Projekt 3D-GUIde für die Evaluation von 3D-CUI-Interaktionslösungen eingesetzt werden kann. Auf dieser Basis kann eine neue formative und summative Evaluationsmethode namens IntuiBeat (d.h. formative Evaluation: IntuiBeat-F, summative Evaluation: IntuiBeat-S) entwickelt werden, was nachfolgend im vierten Kapitel dieser Arbeit vorgestellt wird.

4 Evaluation intuitiver Benutzung mit IntuiBeat

In den vorigen drei Kapiteln wurde erklärt, (1) warum zeitlich effizient einsetzbare formative und summative Evaluationsmethoden im Anwenderprojekt 3D-GUIde benötigt werden, (2) wie intuitive Benutzung im Rahmen dieser Arbeit definiert und auf Basis der vorgestellten Arbeitsdefinition (als Messdefinition intuitiver Benutzung) evaluiert werden kann, und (3) warum sich aktuelle formative und summative Evaluationsmethoden aufgrund ihres zeitlichen Effizienzdefizits nicht als Benchmarks für die Evaluation von User Interfaces (speziell 3D-CUI-Interaktionslösungen im Rahmen des Anwenderprojekts) eignen. Dieses Kapitel beschäftigt sich nun damit, inwiefern dieses Effizienzdefizit durch die neue Evaluationsmethode IntuiBeat behoben werden kann. Hierzu werden aus einer summativen und formativen Perspektive die Grundlagen von IntuiBeat vorgestellt (d.h. summative Evaluation: IntuiBeat-S; formative Evaluation: IntuiBeat-F), die wissenschaftliche Güte von IntuiBeat auf Basis empirischer Befunde diskutiert und damit verbundene Limitationen aufgezeigt. Auf Basis dieser Limitationen werden im Anschluss die Forschungsfragen dieser Arbeit konkretisiert und schließlich ein Überblick über die im Rahmen dieser Arbeit durchgeführten sieben Experimente gegeben, um diese Forschungsfragen zu beantworten.

4.1 Summative Evaluation mit IntuiBeat-S

4.1.1 Empirische Befunde zur Güte der summativen Evaluation

Im Bereich der Lernforschung, in dem wie bei CUIs ebenfalls hauptsächlich Desktop-Systeme beim E-Learning eingesetzt werden (Clark & Mayer, 2016), sind Wissenschaftler ebenfalls mit der Auswahl einer geeigneten universellen Zweitaufgabe konfrontiert, die die nötige Interferenz mit der Hauptaufgabe (d.h. Aufgabe in einer Lernumgebung) bei gleichzeitig geringer Intrusion erzeugt, um eine valide Aussage zur gesamten mentalen Beanspruchung beim Lernenden treffen zu können (Korbach et al., 2018; Park & Brünken, 2015). In der Lernforschung geht man basierend auf der Cognitive Load Theory (Sweller, 1994; Sweller, Ayres, & Kalyuga, 2011) davon aus, dass sich die komplette mentale Arbeitsbelastung und damit auch die gemessene mentale Beanspruchung (d.h. mentale Effizienz) beim Lernen aus drei Komponenten ergibt: (1) intrinsische mentale Arbeitsbelastung (d.h. mentale Arbeitsbelastung, die durch die Komplexität des Lernmaterials selbst bedingt ist), (2) extrinsische mentale Arbeitsbelastung (d.h. mentale Arbeitsbelastung, die durch die Darstellung und Gestaltung des Lernmaterials bedingt ist) und (3) lernbezogene mentale Arbeitsbelastung (d.h. mentale Arbeitsbelastung, die beim Lernenden aufgewendet wird, um das Lernmaterial zu verstehen und damit verbundene Schemata zu verarbeiten, zu aktivieren und zu automatisieren) (Korbach et al., 2018; Park & Brünken, 2015). Obwohl sich die gesamte mentale Arbeitsbelastung beim Lernen somit unterschiedlich zusammensetzt, wird, wie in der HCI, auch in der Lernforschung versucht, die gesamte mentale

Beanspruchung des Arbeitsgedächtnisses bei der summativen Evaluation zu erfassen. Daher wird diese Aufteilung im weiteren Verlauf dieser Arbeit ignoriert, wurde aber dennoch erwähnt, um den theoretischen Hintergrund der folgenden empirischen Befunde aus der Lernforschung zu kennen.

Zweitaufgaben im Bereich des E-Learnings sind üblicherweise akustisch oder visuell dargebotene Hinweisreize, die in die Lernumgebung eingebettet werden. Beispielsweise mussten in einer Studie von Brünken, Plass und Leutner (2004) Teilnehmer während des Lernens einen Buchstaben im oberen Bereich des Bildschirms beobachten und die Leertaste drücken, sobald dieser die Farbe änderte. Hier lässt sich die gleiche Problematik aufgrund der im letzten Kapitel angesprochenen Strukturwechseleffekts beobachten (Park & Brünken, 2015). Beide Aufgaben werden in der gleichen Modalität (d.h. visuell) dargeboten und können dadurch die gesamte modalitätsspezifische mentale Beanspruchung abbilden, laufen aber gleichzeitig Gefahr, die wissenschaftliche Güte der Messung durch eine hohe Intrusivität der Zweitaufgabe zu gefährden. Eine Inhibitionsaufgabe als Zweitaufgabe könnte laut Park und Brünken (2015) eine hohe Interferenz erzeugen, ohne dabei besonders intrusiv zu sein, wenn diese modalitätsunabhängig gegenüber der Hauptaufgabe ist und identische kognitive Ressourcen wie die Hauptaufgabe nutzt. Wie bereits im letzten Kapitel erwähnt, gelten Inhibitionsprozesse als gute Indikatoren für Prozesse der zentralen Exekutive (J. D. Cohen et al., 1997; A. Miyake et al., 2000) und können demnach als universelle Indikatoren mentaler Arbeitsbelastung betrachtet sowie zur Evaluation mentaler Beanspruchung genutzt werden (Park & Brünken, 2015).

Park und Brünken (2015) entwickelten auf Basis dieser Erkenntnisse ihre „Rhythmus-Methode“, eine Zweitaufgabe, die explizit Inhibitionsprozesse anspricht. Diese Methode bildet die Grundlage von IntuiBeat, einer Adaption dieser Methode für den HCI-Bereich zur Evaluation intuitiver Benutzung. Sie verlangt von Nutzern, dass sie einen zuvor erlernten einfachen Rhythmus im Viervierteltakt mit Pausen (d.h. Schlag-Schlag-Pause-Pause/Schlag-Schlag-Pause-Pause) mit dem Fuß klopfen müssen. Wie in Abbildung 4.1 zu erkennen ist, besteht ein solcher Rhythmus aus zwei Komponenten: einer *kurzen Rhythmuskomponente* und einer *langen Rhythmuskomponente*. Die kurze Rhythmuskomponente umfasst 500 Millisekunden und beschreibt das Intervall zwischen den beiden Schlägen. Im Gegenzug beschreibt die lange Komponente die Dauer der beiden Pausen bzw. den Abstand zum Beginn der nächsten Rhythmuseinheit. Die lange Komponente umfasst 1500 Millisekunden. Da die Nutzer dabei kontinuierliches Klopfen unterdrücken und stattdessen bewusst Pausen einhalten müssen (d.h. kognitives Abkoppeln), ist ihre zentrale Exekutive belastet und Änderungen in der mentalen Arbeitsbeanspruchung einer parallel durchgeführten Hauptaufgabe werden in den Rhythmusabweichungen sichtbar. Rhythmusabweichungen können daher eine gewisse Inhaltsvalidität zugesprochen werden (Korbach et al., 2018; Park & Brünken, 2015).

Da erst die bewusste Pause eine Inhibition ermöglicht und somit die Erfassung mentaler Effizienz gestattet, erlaubt die von Park und Brünken (2015) vorgeschlagene etwas längere Pause (d.h. 1500 Millisekunden) mit hoher Wahrscheinlichkeit diesen Inhibitionsprozess präzise zu erfassen. Die Rhythmusabweichungen berechnen sich schließlich aus einer zuvor ermittelten individuellen Baseline. Auf diese Weise berücksichtigt die Methode interindividuelle Unterschiede und erlaubt dadurch eine summative Evaluation mentaler Effizienz. Die aufgrund des ausgewählten Rhythmus eingehandelte geringe zeitliche Auflösung bzw.

die nicht vorhandene Echtzeitfähigkeit (d.h. Zweitaufgabe und Erstaufgaben können nicht exakt parallel ausgeführt werden) ist generell ein Problem bei der Verwendung von Zweitaufgabenleistungsmaßen (z.B. Erfassung von Reaktionszeiten bei der Generierung von Zufallsnummern) und nicht nur speziell bei der Rhythmusmethode. Es sollte jedoch auf Basis empirischer Ergebnisse von anderen Zweitaufgaben (siehe Teilabschnitt 3.6.2.4) bei einer summativen Evaluation nicht groß ins Gewicht fallen, da es sich ja dabei um die zusammenfassende Bewertung eines Evaluationsgegenstandes handelt und nicht um die formative Bewertung einzelner Ereignisse und Nutzungsprobleme.

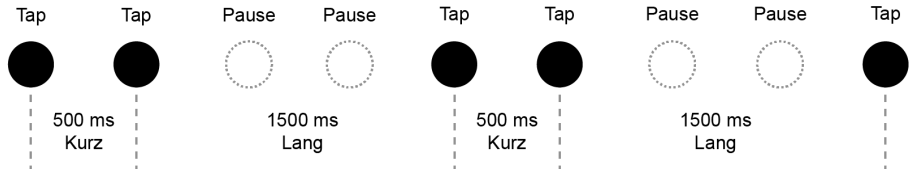


Abbildung 4.1. Der bei der Rhythmusmethode zu klopfende Rhythmus im Viervierteltakt (Park & Brünken, 2015). Eine Rhythmuseinheit besteht immer aus einem kurzen Rhythmusintervall von 500 Millisekunden (schwarz) und einem langen Rhythmusintervall von 1500 Millisekunden (weiß).

Aufgrund der direkten Assoziation mit der zentralen Exekutive besitzt eine solche Rhythmusaufgabe eine hohe Interferenz gegenüber verschiedenen Arten von Hauptaufgaben und erfordert nicht das bewusste Unterbrechen derselbigen wie es beispielsweise bei der Stroop-Aufgabe von Gwizdka (2010) der Fall war, da die Rhythmusaufgabe selbstbestimmt ohne externe Hinweisreize auskommt (Park & Brünken, 2015). Dementsprechend erfüllt die Methode alle von Cain (2007) formulierten Anforderungen an Zweitaufgaben: leichte Erlernbarkeit, hohe Selbstbestimmtheit und die Nichtbeeinflussung der Hauptaufgabe durch unerwünschte Unterbrechungen (d.h. geringe Intrusion). Sie besitzt somit auch hohes Potential als universelle Zweitaufgabe für die summative Evaluation intuitiver Benutzung von User Interfaces (speziell 3D-CUI-Interaktionslösungen im Rahmen des Projekts 3D-GUIde).

Park und Brünken (2015) nutzten die Rhythmusmethode zur Evaluation von zwei, im Bereich des E-Learnings sehr verbreiteten, Effekten des Intruktionsdesigns: des *Modalitätseffekts* (d.h. modality effect) und des *Effekts der verlockenden Details* (d.h. seductive details effect). Der Modalitätseffekt beschreibt die lernförderliche Wirkung, die durch die gemeinsame Nutzung der akustischen und visuellen Modalität des Arbeitsgedächtnisses entsteht (Sweller, 1994). Als Effekt der verlockenden Details wird der lernhinderliche Effekt durch das Hinzufügen interessanter, aber lernirrelevanter Zusätze bezeichnet (Garner, Gillingham, & White, 1989; Harp & Mayer, 1998; Lehman, Schraw, McCrudden, & Hartley, 2007). Beide Effekte beeinflussen nachweislich die mentale Arbeitsbelastung und damit auch die messbare mentale Beanspruchung während des Lernvorgangs (Garner et al., 1989; Harp & Mayer, 1998). Im ersten Experiment ($N = 30$) variierten Park und Brünken (2015) die verlockenden Details des Lernmaterials, das aus verschiedenen Bildern und Texten über die ATP-Synthase bestand (hohe mentale Arbeitsbelastung: mit verlockenden Details; geringe mentale Arbeitsbelastung: ohne verlockende Details). Beim zweiten Experiment ($N = 50$) nutzten die Autoren dasselbe Lernmaterial, manipulierten jedoch nicht die verlo-

ckenden Details, sondern die Modalität des Lernmaterials (hohe mentale Arbeitsbelastung: nur Lernmaterial, d.h. Bildschirmtexte und Bilder; niedrige mentale Arbeitsbelastung: Bilder und Bildschirmtexte ersetzt durch Audiokommentar eines männlichen Sprechers). In beiden Experimenten kam ein Between-Subjects Design zum Einsatz.

In beiden Experimenten mussten die Lernenden zuvor den in Abbildung 4.1 dargestellten Rhythmus lernen. Ihnen wurde dazu zunächst der Rhythmus einmal über Lautsprecher in Form einer Audiodatei vorgespielt. Im Anschluss wurde ihnen der Rhythmus erneut vorgespielt, wobei sie dieses Mal parallel zur Audioaufzeichnung klopfen mussten. Danach mussten sie den Rhythmus alleine ohne Unterstützung der Audioaufzeichnung eine Minute lang klopfen. Diese letzte Minute wurde mithilfe einer Aufnahmesoftware aufgezeichnet und fungierte für den Probanden als individuelle Baseline für die gesamte Lernsitzung. Laut Park und Brünken (2015) ist die Erfassung einer Baseline sinnvoll, da die Messungen so unabhängig von individuellen Unterschieden im Rhythmusgefühl und den Vorerfahrungen mit der Methode sind. Gleichzeitig ist es durch eine Baseline möglich, intraindividuelle Vergleiche der erfassten Rhythmusabweichungen unter verschiedenen experimentellen Bedingungen vornehmen zu können. Damit wird gewährleistet, die mentale Effizienz von Probanden auch in Between-Subjects Designs basierend auf Einzelwertungen beurteilen zu können, da die Einzelwertungen jedes Probanden immer in Zusammenhang mit der individuellen Baseline seiner mentalen Beanspruchung stehen. Nach der Erhebung der Baseline führten die Probanden die eigentliche Lernaufgabe durch. Sie wurden hier instruiert, während der Instruktion des Lernmaterials und dem eigentlichen Lernen den gelernten Rhythmus weiter zu klopfen.

Wie zu erwarten, zeigte sich in beiden Experimenten (d.h. Experiment 1: mit verlockenden Details; Experiment 2: nur Bildschirmtexte ohne zusätzlichen Audiokommentar des Bildmaterials) eine signifikant höhere mittlere Rhythmusabweichung bezüglich der individuellen Baseline bzw. eine geringere Rhythmuspräzision bei höherer mentaler Arbeitsbelastung gegenüber der jeweiligen Kontrollgruppe. Dieser Zusammenhang zeigte sich jedoch nur für die kurze Rhythmuskomponente. Eine Erklärung sehen Park und Brünken (2015) in dem Phänomen, das oft auch bei ereigniskorrelierten Potentialen im Gehirn auftritt, in dem sich Inhibitionsprozesse erst zeitversetzt und nicht sofort zeitgleich mit dem eigentlichen Inhibitionsereignis zeigen (siehe Berti, 2008). Müssen Probanden ihr regelmäßiges Klopfen aufgrund einer Pause bewusst unterbrechen (d.h. lange Rhythmuskomponente), wirkt sich diese Unterbrechung erst auf die nachfolgende Aktion aus (d.h. Auswirkung auf die Schlagfrequenz der kurzen Rhythmuskomponente, siehe Abbildung 4.1) (Park & Brünken, 2015). Im weiteren Verlauf der Arbeit werden deswegen lediglich die Rhythmusabweichungen der kurzen Rhythmuskomponente als Indikator für mentale Arbeitsbelastung und damit für die mentale Effizienz der kognitiven Informationsverarbeitung betrachtet.

Um erste Aussagen über die wissenschaftliche Güte der Rhythmusmethode als objektives Maß für die mentale Effizienz während der kognitiven Informationsverarbeitung treffen zu können, ließen Park und Brünken (2015) die Probanden am Ende jeder Lernsitzung mithilfe eines Fragebogens ihre mentale Effizienz subjektiv bewerten. Damit erfassten sie ein subjektives Quasi-Außenkriterium für die Meta-Evaluation ihrer Methode. Sie nutzten dafür die Skala zur Erfassung der investierten mentalen Anstrengung (Subjective Rating Scale), die im Lernbereich ein übliches Maß darstellt (Paas, 1992; Paas, Tuovinen, Tabbers, & Van Gerven, 2003; Paas & Van Merriënboer, 1994). Diese Skala besteht aus zwei Items:

eines zur Erfassung der mentalen Beanspruchung zum Verständnis des Lernmaterials und eines zur Erfassung der mentalen Beanspruchung zum Bearbeiten des Lernmaterials. Beide Items (d.h. „Wie leicht oder schwer ist das Lernmaterial zu verstehen?“, „Wie hoch oder niedrig ist ihre mentale Anstrengung bei der Bearbeitung des Lernmaterials?“) werden dazu auf einer siebenstufigen Likertskala von „sehr niedrig/leicht“, „niedrig/leicht“, „ziemlich niedrig/leicht“, „weder niedrig/leicht noch hoch/schwierig“, „ziemlich hoch/schwierig“, „hoch/schwierig“ bis „sehr hoch/schwierig“ retrospektiv bewertet (Korbach et al., 2018; Paas, 1992; Paas & Van Merriënboer, 1994; Park & Brünken, 2015).

Im Vergleich zu den in Abschnitt 3.6.1 vorgestellten Fragebögen SEA und NASA-TLX stellt die Skala jedoch aufgrund ihrer nur siebenstufigen Skala ein sehr grobes Maß dar. In diesem Zusammenhang veröffentlichte aktuelle Meta-Analysen sprechen dem Maß zwar eine hohe Reliabilität zu (Cronbachs $\alpha > .70$ auf Basis einer Konsistenzanalyse; Hadie & Yusoff, 2016), einige Meta-Analysen stellen jedoch auch die Valididät der Skala in Frage (z.B. Moreno, 2010), obwohl verschiedene empirische Befunde sogar die Konstruktvalidität der Skala zeigen konnten (z.B. Ayres, 2006; Paas et al., 2003; Szulewski, Gegenfurtner, Howes, Sivilotti, & Van Merriënboer, 2017). Naismith und Cavalcanti (2015) merken in ihrer in diesem Zusammenhang durchgeführten Meta-Analyse beispielsweise an, dass der Fragebogen in vielen Studien in abgeänderter Form (d.h. andere Formulierungen der Items, und Änderung der Abstufung der Likertskala) dargeboten wird, deren Validität jedoch noch nicht explizit sichergestellt wurde.

Park und Brünken (2015) verwendeten den Fragebogen in einer Form mit einer siebenstufigen Likertskala, wobei die Konstruktvalidtät lediglich für die neunstufige Version nachgewiesen wurde (siehe Ayres, 2006; Paas et al., 2003; Szulewski et al., 2017). Zusätzlich nutzten sie auch nur das Item zur Messung der mentalen Beanspruchung, da sie annahmen, dass dies auch die mentale Beanspruchung beim Verständnis impliziere. Darüber hinaus erfassten sie den allgemeinen Lernerfolg mithilfe eines Fragebogens als Hauptaufgabenleistungsmaß, den sie aber nicht als Quasi-Außenkriterium bei ihrer Meta-Evaluation zur Sicherstellung der wissenschaftlichen Güte der Rhythmusmethode berücksichtigten. In beiden Experimenten konnte kein statistisch signifikanter Zusammenhang zwischen den mittleren Rhythmusabweichungen und der subjektiven Bewertung mentaler Beanspruchung festgestellt werden (Experiment 1: $r(28) = .02$, n. s.; Experiment 2: $r(45) = .13$, n. s.). Daher ist an dieser Stelle die Kriteriumsvalidität der Methode erstmal anzuzweifeln. Die Reliabilität der Rhythmusmethode, bestimmt mittels Testhalbierungsreliabilität nach Guttman (1945), war hingegen sowohl im ersten Experiment ($r = .96$) als auch im zweiten Experiment ($r = .78$) hoch, wenn man die Benchmarks von Field (2017) zugrunde legt. Die Objektivität der Rhythmusmethode wurde von Park und Brünken (2015) nicht explizit bei der Meta-Evaluation der Rhythmusmethode thematisiert.

Park und Brünken (2015) erklärten ihre nicht statistisch signifikanten Ergebnisse damit, dass durch die objektive Rhythmuszweitaufgabe und den subjektiven Fragebogen unterschiedliche Aspekte mentaler Beanspruchung gemessen werden. Die Rhythmusaufgabe bildet die tatsächlich benötigte gesamte mentale Arbeitsbelastung anhand von exekutiven Inhibitionsprozessen ab (siehe A. Miyake et al., 2000). Die subjektiv wahrgenommene mentale Beanspruchung kann sich davon jedoch unterscheiden. Wie bereits in Kapitel 3.6.1 angesprochen, können retrospektiv erhobene subjektive Maße unter anderem durch Biases wie Primacy-Recency-Effekte verzerrt werden, für die ein objektives, kontinuierliches Maß

weniger anfällig ist (siehe Tretter et al., 2018). Daher sollten für eine Meta-Evaluation nicht nur subjektive Quasi-Außenkriterien genutzt werden. Sich dabei wie Park und Brünken (2015) nur auf ein Quasi-Außenkriterium zu verlassen und davon nur die Hälfte an Items zu verwenden, ist aus methodischer Sicht dementsprechend fragwürdig. Zusätzlich besitzt die von Paas (1992) vorgeschlagene Skala zur Erfassung der investierten mentalen Anstrengung laut des bereits angeführten Reviews von Naismith und Cavalcanti (2015), im Vergleich zu Messinstrumenten wie dem NASA-TLX, eine geringere Validität.

Die Vorgehensweise bei der Meta-Evaluation kann daher den Befund von Park und Brünken (2015) erklären und die Daten des Experiments liefern auch einen vorläufigen Hinweis darauf. Da sich der Aufbau der Lernaufgabe (d.h. gleiche Anzahl von Folien; gleiche Instruktionen und Lerninhalte) über die Bedingungen hinweg (d.h. hohe vs. geringe mentale Arbeitsbelastung) nicht geändert hat, könnte eine besonders schwere Folie zum Schluss die subjektiv wahrgenommene mentale Beanspruchung systematisch verzerrt haben. Dies ist laut Park und Brünken (2015) wahrscheinlich, da sich keine signifikanten Unterschiede bezüglich der subjektiv wahrgenommenen mentalen Beanspruchung zwischen den Bedingungen feststellen ließen. Dies impliziert, dass für die Lernenden alle Bedingungen subjektiv gleich komplex waren, was aufgrund der gewählten sehr starken Manipulation erstaunlich ist (Experiment 1: $F(1,26) < 1$, n. s., d = 0.04; Experiment 2: $F(1,43) = 1.31$, n. s., d = .34). Jedoch konnte ein statistisch signifikanter Unterschied zwischen der jeweiligen Manipulationsbedingung und Kontrollbedingung bezüglich der mittleren Rhythmusabweichungen festgestellt werden (Experiment 1: $F(1,26) = 5.48$, $p < .05$, d = 1.15; Experiment 2: $F(1,43) = 6.12$, $p < .05$, d = 1.01). Dieser Aspekt macht eine mögliche Verzerrung des subjektiven Maßes ebenfalls sehr wahrscheinlich. Unabhängig von den Ergebnissen von Park und Brünken (2015) konnte objektiven Maßen im Bereich des E-Learnings generell eine höhere Sensitivität bei der Evaluation mentaler Beanspruchung im Vergleich zu subjektiven Maßen nachgewiesen werden (siehe DeLeeuw & Mayer, 2008). Trotzdem kann auf Basis dieser Ergebnisse noch nicht mit Sicherheit gesagt werden, ob die Rhythmusmethode wirklich mentale Beanspruchung oder ein anderes Merkmal misst.

Da Park und Brünken (2015) in ihren beiden Experimenten lediglich einen ersten Hinweis für die wissenschaftliche Tragfähigkeit von Rhythmusabweichungen zur Evaluation mentaler Beanspruchung anhand einer uneindeutigen Meta-Evaluation, aber keinen vollständigen empirischen Nachweis der Validität liefern konnten, evaluierten sie das Maß anhand von vier weiteren objektiven Quasi-Außenkriterien und einem subjektiven Quasi-Außenkriterium in einem Folgeexperiment erneut. Als Studiendesign kam beim Folgeexperiment von Korbach et al. (2018) ein Between-Subjects Design mit drei Gruppen (n_1 = 26, n_2 = 26, n_3 = 26) und das gleiche Lernmaterial wie bei der Originalstudie (siehe Park & Brünken, 2015) zum Einsatz. Bei der ersten Gruppe wurde das Lernmaterial wie in der Originalstudie zusätzlich mit verlockenden Details (d.h. irrelevante Informationen) angereichert, um die mentale Arbeitsbelastung der Nutzer durch erhöhte Ablenkung vom eigentlichen Lernprozess zu erhöhen (siehe Garner et al., 1989; Rey, 2012, 2014). Diese Bedingung stellt somit eine Replikation des ersten Experiments dar. Die zweite Gruppe von Probanden erhielt ebenfalls das gleiche Lernmaterial, das aber dieses Mal mit dreidimensionalen Animationen angereichert wurde. Diese Animationen unterstützten den Lernprozess zwar, indem sie dem Lernenden halfen, sich mentale dreidimensionale Repräsentationen aus dem statischen zweidimensionalen Lernmaterial (d.h. Bilder und Text) vorzustellen, erhöhten aber so gleichzeitig die mentale Arbeitsbelastung (Korbach et al., 2018). Der

dritten Gruppe wurde das Lernmaterial ohne jegliche Manipulation präsentiert. Sie stellt somit eine Kontrollgruppe dar. Die Reliabilität der Rhythmusabweichungen wurde wie in der Originalstudie mittels Testhalbierungsreliabilität nach Guttman (1945) bestimmt und war mit $r = .96$ exzellent, sofern man die Benchmarks von Field (2017) zugrunde liegt. Wie bereits in der Originalstudie wurde Objektivität von den Forschern nicht explizit bei der Meta-Evaluation der Rhythmusmethode thematisiert und sich stattdessen neben der Reliabilität auf die Validität konzentriert.

Zur Sicherstellung der Validität wählten Korbach et al. (2018) die Kriteriumsvalidität als methodischen Zugang und verglichen dazu die drei beschriebenen Gruppen hinsichtlich aufgezeichneter mittlerer Rhythmusabweichungen der kurzen Rhythmuskomponente, des in der Originalstudie ebenfalls verwendeten Fragebogens (Paas, 1992) und der drei objektiven Maße für Augenbewegungen: Index of Cognitive Activity (ICA) (S. P. Marshall, 2002), der Anzahl der Übergänge zwischen Textbereichen (sog. AOIs: Area of Interest) und Bildbereichen (Schmidt-Weigand, Kohnert, & Glowalla, 2010), sowie damit verbundenen Fixationszeiten. Wie bereits in Abschnitt 3.6 beschrieben, wird der ICA auf Basis von Blickdaten, die mithilfe eines Eyetrackers aufgezeichnet werden, berechnet. Der ICA bildet dabei die mentale Beanspruchung des Nutzers durch Analyse der Kurzfristig- und Unregelmäßigkeit des psychosensorischen Reflexes der Pupille ab (S. P. Marshall, 2002). Eine Reihe von Studien konnte bereits belegen, dass neben ICA auch die Anzahl von Fixationen, die Dauer der Fixationen und die Übergänge zwischen Text- und Bildbereichen als Indikatoren für mentale Arbeitsbelastung und damit zur Evaluation mentaler Beanspruchung genutzt werden können. Sie können darüber hinaus auch Auskunft über die Ursachen für die erhöhte mentale Beanspruchung geben (Korbach et al., 2018; Park & Brünken, 2015).

Aufgrund der nicht nachgewiesenen Kriteriumsvalidität in der Originalstudie (siehe Park & Brünken, 2015) fragten die Autoren in der Folgestudie sicherheitshalber beide Items ihres subjektiven Quasi-Außenkriteriums ab (d.h. in der Mitte des Lernvorgangs und am Ende des Lernvorgangs) und berücksichtigten beide Items auch in ihrer Meta-Evaluation. Wie bereits in der Originalstudie verwendeten sie jedoch weiterhin eine siebenstufige Likertskala anstatt einer neunstufigen, wie es beim Fragebogen von Paas (1992) normalerweise üblich ist (siehe Ayres, 2006; Paas et al., 2003; Szulewski et al., 2017). Des Weiteren wurde das subjektive Lernverständnis mit einem Fragebogen gemessen, der aus acht Items bestand und eine gute interne Konsistenz (Cronbachs $\alpha = .71$) zeigte. Das Lernverständnis fungierte für die Autoren als zusätzliches objektives Quasi-Außenkriterium, welches mit einem Lerntest abgefragt wurde. Es wurden ferner noch weitere Maße zur Bestimmung des Lernerfolgs durch die Autoren erhoben, die die Transferierbarkeit und das Behalten des gelernten Wissens nachweisen sollten. Aufgrund der Tatsache, dass das Verständnis des Lernmaterials notwendig für seine Transferierbarkeit auf andere Situationen und sein dauerhaftes Beibehalten ist (siehe Royer, 1979), werden im Folgenden einfachheitshalber nur die Ergebnisse bezüglich des Lernverständnisses als Quasi-Außenkriterium für den Lernerfolg berichtet. Darüber hinaus wird außerdem nur auf die Ergebnisse des Fragebogens eingegangen, der am Ende des Lernvorgangs erhoben wurde. Für die anderen Maße und ihre genaue Operationalisierung wird auf die Veröffentlichung von Korbach et al. (2018) verwiesen, da die Ergebnisse nahezu identisch sind.

Eine Varianzanalyse bezüglich des Fragebogens zeigte, auch wenn der Fragebogen mit beiden Items verwendet wurde, keine statistisch signifikanten Unterschiede zwischen den drei Gruppen (Item „Schwierigkeit des Lernmaterials“: $F(2, 74) = 2.486$, n. s., Item „Wahrgenommene mentale Arbeitsbelastung“: $F(2, 74) = 3.108$, n. s.). Bezüglich des ICA konnten ebenfalls keine signifikanten Unterschiede zwischen den Gruppen festgestellt werden, $F(2,74) < 1$, n. s. An dieser Stelle soll darauf hingewiesen werden, dass Korbach et al. (2018) keine Effektgrößen, Mittelwerte oder Standardabweichungen veröffentlicht haben und deswegen auch keine Effektgrößen nachträglich berechnet werden können. Bezüglich der mittleren Rhythmusabweichungen zeichnete sich erwartungsgemäß ein signifikanter Unterschied zwischen den Gruppen ab, $F(2, 74) = 3.197$, $p < .05$, d = .59. Eine Varianzanalyse bezüglich der Augenbewegungen zeigte signifikante Effekte der Gesamtfixationszeit auf Bildflächen ($F(2, 74) = 6.682$, p < .05, d = .85), der Gesamtfixationszeit auf Textflächen ($F(2, 74) = 8.083$, $p < .05$, d = .93) und der Gesamtanzahl an Übergängen zwischen Bild- und Textflächen ($F(2, 74) = 4.671$, $p < .05$, d = .71). Für die Analyse der geplanten Kontraste und der damit verbundenen Teststatistiken der eben beschriebenen Maße wird auf Korbach et al. (2018) verwiesen, da diese Informationen für die Beurteilung der wissenschaftlichen Güte von Rhythmusabweichungen als Maß für die mentale Effizienz bei der kognitiven Informationsverarbeitung nicht unbedingt notwendig sind.

Zur Überprüfung der Kriteriumsvalidität wurden von Korbach et al. (2018) zunächst Korrelationen zwischen allen erhobenen objektiven Maßen berechnet. Die mittleren Rhythmusabweichungen zeigten dabei wider Erwarten keine signifikanten Korrelationen mit den erhobenen objektiven Augenmaßen, wobei fast alle objektiven Augenmaße (Ausnahme: keine Korrelation zwischen dem ICA und der Gesamtfixationszeit auf Bildflächen) untereinander signifikante Korrelationen zeigten. Alle erhobenen objektiven Maße, bis auf der ICA und die Gesamtanzahl an Übergängen zwischen Bild- und Textflächen, korrelierten signifikant mit dem Lernerfolg. Dies spricht laut Korbach et al. (2018) für eine vorhandene Sensibilität der Maße gegenüber mentaler Beanspruchung. In Übereinstimmung mit den Erkenntnissen aus der Originalstudie von Park und Brünken (2015) zeigten die mittleren Rhythmusabweichungen auch in der Folgestudie keine signifikante Korrelation mit dem ersten Item der Subjective Rating Scale ($r(78) = .011$; n. s.). Die anderen objektiven Maße konnten hier allerdings auch keinen signifikanten Zusammenhang mit der mentalen Beanspruchung feststellen. Bezüglich des zweiten Items der Subjective Rating Scale, ließ sich allerdings eine signifikante Korrelation mit den mittleren Rhythmusabweichungen feststellen, $r(78) = .24$, $p < .05$. Wie bereits beim ersten Item konnte auch beim zweiten Item kein signifikanter Zusammenhang zwischen diesem und den anderen objektiven Maßen festgestellt werden. Für die Werte dieser Korrelationen und weitere Informationen wird auf Korbach et al. (2018) verwiesen.

Zusammenfassend zeigten die Varianzanalyse und die berechneten Korrelationen, dass von den erhobenen Maßen lediglich die Fixationszeiten (d.h. Bild und Text) und die Rhythmusabweichungen sensibel genug waren, um die unterschiedliche mentale Arbeitsbelastung in den Gruppen abbilden zu können, und somit auch über eine entsprechende Kriteriumsvalidität verfügen. Sowohl Fixationszeiten als auch die Rhythmusabweichungen bildeten dabei das objektive Quasi-Außenkriterium Lernerfolg ab. Beide Maße scheiterten jedoch daran, die subjektiv wahrgenommene mentale Beanspruchung zu reflektieren. Lediglich die Rhythmusabweichungen spiegelten diese immerhin bezüglich des zweiten Items des Fragebogens wider. Fixationszeiten und Rhythmusabweichungen wiesen dabei keinen Zu-

sammenhang auf. Korbach et al. (2018) postulieren deswegen, dass nicht notwendigerweise ein linearer Zusammenhang zwischen Augenmaßen, Rhythmusabweichungen und subjektiven Bewertungen bestehen muss. So können zwar längere Fixationszeiten auf Bildern den Lernerfolg bzw. das Verständnis der Inhalte verbessern, müssen sich aber nicht gleichzeitig unmittelbar in der mentalen Beanspruchung des Lernenden zeigen. Die Analyse von Fixationszeiten und die damit verbundenen Bildflächen können aber in Zusammenhang mit der Rhythmusmethode die Kausalitäten aufzeigen, die für die Erhöhung der mentalen Beanspruchung verantwortlich sind. Wie bereits in Abschnitt 3.6 angesprochen, wird die wissenschaftliche Güte von okularen Parametern sehr stark von den Beleuchtungsverhältnissen beeinflusst (Cain, 2007; Charles & Nixon, 2019; F. Chen et al., 2016; Cowley et al., 2016), was von Korbach et al. (2018) nicht thematisiert und somit auch nicht kontrolliert wurde.

4.1.2 Diskussion von Limitationen und Bewältigung durch IntuiBeat-S

Auf Basis der uneindeutigen Ergebnisse der Original- und Folgestudie kann der Rhythmusmethode streng genommen noch keine Kriteriumsvalidität als summative Evaluationsmethode für intuitive Benutzung attestiert werden. Dies ist jedoch auch auf die Auswahl und die wissenschaftliche Güte, der von den Autoren verwendeten Quasi-Außenkriterien zurückzuführen. Die Verwendung von überwiegend okularen physiologischen Maßen, ohne explizit die Beleuchtungsverhältnisse zu kontrollieren (siehe Cain, 2007; Charles & Nixon, 2019; F. Chen et al., 2016; Cowley et al., 2016) stellt eine große Einschränkung der Folgestudie von Korbach et al. (2018) dar. Diese Einschränkung lässt an der wissenschaftlichen Güte dieser verwendeten objektiven Quasi-Außenkriterien zweifeln. In beiden Studien kam mit der Skala zur Erfassung der investierten mentalen Anstrengung von Paas (1992) im Vergleich zu den im HCI-Bereich zur Erfassung intuitiver Benutzung gängigen Fragebögen SEA-Skala und NASA-TLX ein recht grobes subjektives Maß als Quasi-Außenkriterium zum Einsatz (z.B. Verwendung einer neunstufigen Likertskala ist gegenüber einem zwanzigstufigen NASA-TLX als grob einzustufen). Die wissenschaftliche Güte dieser Skala ist zwar durch einige Studien gesichert, aber nur, wenn dieser auch in der vorgesehenen Art und Weise administriert wird.

Da in beiden Originalstudien zur Rhythmusmethode anstelle einer neunstufigen eine siebenstufige Likertskala zum Einsatz kam und in der Originalstudie nur ein Item des Fragebogens für die subjektive Erfassung mentaler Beanspruchung hinzugezogen wurde, lässt sich die Qualifikation dieses Maßes als subjektives Quasi-Außenkriterium anzweifeln. Dennoch konnte die vorgestellte Folgestudie im Gegensatz zur Originalstudie immerhin zeigen, dass mittlere Rhythmusabweichungen ähnlich gut die mentale Beanspruchung wie objektive Hauptaufgabenleistungsmaße (d.h. Effektivität, ein definierendes Merkmal intuitiver Benutzung im Rahmen der neuen Arbeitsdefinition, siehe Abschnitt 2.2) und subjektive Fragebögen widerspiegeln können. Allerdings muss in zukünftigen Studien unbedingt untersucht werden, ob die Musikalität der Versuchsteilnehmer Einfluss auf die Rhythmusabweichungen der Probanden hat. Zudem fanden die vorgestellten Studien noch nicht im CUI-Bereich statt.

Da der noch ausstehende Nachweis der wissenschaftlichen Güte der Rhythmusmethode somit wahrscheinlich auf die Vorgehensweise bei der Meta-Evaluation in den Original-

studien zurückzuführen ist und die Rhythmusmethode an sich eine hohe Inhaltsvalidität wegen ihrer Verbindung zu Inhibitionsprozessen, sowie der damit verbundenen Erfassung mentaler Effizienz anhand exekutiver Funktionen besitzt, bietet sie ein enormes Potential für die summative Evaluation intuitiver Benutzung im HCI-Bereich. Um jedoch Rhythmusabweichungen für die summative Evaluation von User Interfaces (speziell 3D-CUI-Interaktionslösungen im Zuge des Anwenderprojekts 3D-GUIde) zeitlich effizient einsetzen zu können, müssen zur Sicherstellung ihrer wissenschaftlichen und zeitlichen Anwendungseffizienz eine Reihe von Anforderungen erfüllt werden:

- Adaption der Rhythmusmethode als IntuiBeat-S im Bereich der Mensch-Maschine-Interaktion, um sie zur summativen, auf mentaler Effizienz basierenden Evaluation intuitiver Benutzung bei User Interfaces (speziell 3D-CUIs im Rahmen des Anwenderprojekts 3D-GUIde) zeitlich effizient einsetzen zu können.
- Sicherstellung der wissenschaftlichen Güte von IntuiBeat-S als summative Evaluationsmethode für intuitive Benutzung bei User Interfaces (speziell 3D-CUIs) anhand der Hauptgütekriterien *Objektivität*, *Validität* und *Reliabilität* auf Basis der in Abschnitt 3.6 identifizierten subjektiven und objektiven Quasi-Außenkriterien.
- Sicherstellung der *zeitlichen Anwendungseffizienz* von IntuiBeat-S durch den Vergleich mit dem in Abschnitt 3.6 identifizierten aktuellen Benchmark zur summativen Evaluation intuitiver Benutzung, der CHAI-Methode.
- Überprüfung etwaiger Konfundierungen der Methode durch Unterschiede in der Musikalität der Systemnutzer.

4.2 Formative Evaluation mit IntuiBeat-F

4.2.1 Empirische Befunde zur Güte der formativen Evaluation

Bei der summativen Evaluation werden Rhythmusabweichungen zur quantitativen Bewertung eines Systems bezüglich des damit verbundenen Ausmaßes an intuitiver Benutzung genutzt. Im Gegensatz dazu werden bei einer formativen Evaluation potentielle, intuitive Benutzung beeinträchtigende Nutzungsprobleme zuvor mithilfe von Rhythmusabweichungen aufgedeckt. Wo bei einer summativen Evaluation die Rhythmusabweichungen im Zusammenhang mit der Baseline des Nutzers direkt zur Quantifizierung intuitiver Benutzung genutzt werden können, muss bei der formativen Evaluation zuvor ein Schwellenwert festgestellt werden, ab dem sich hinter einer Rhythmusabweichung ein Nutzungsproblem verbirgt. Da sich die Arbeiten von Park und Brünken (2015), sowie auch die spätere Arbeit von Korbach et al. (2018), ausschließlich der summativen Evaluation mentaler Beanspruchung widmeten und formative Aspekte keine Rolle spielten, sollten zunächst andere relevante Arbeiten zu Rhythmusaufgaben zur Erfassung mentaler Effizienz im HCI-Bereich betrachtet werden.

So nutzten Tracy und Albers (2006), sowie M. J. Albers (2011) eine Klopfaufgabe, um die mentale Beanspruchung bei der Nutzung von Webseiten zu evaluieren. Obwohl die Forscher im Vergleich zu den Arbeiten aus dem Lernbereich (siehe Korbach et al., 2018; Park & Brünken, 2015) unterschiedliche theoretische Grundlagen nutzten, lassen sich bei

genauerer Betrachtung dennoch viele Überschneidungen der Arbeiten erkennen. Wo Park und Brünken (2015) ihre Rhythmusmethode eher allgemein auf Inhibitionsprozessen fundierten, stützen sich M. J. Albers (2011) auf Y. Miyake et al. (2004), die rhythmisches Klopfen direkt mit der zentralen Exekutive des Arbeitsgedächtnisses in Verbindung bringen. Diese Verbindung zur zentralen Exekutive wird jedoch von M. J. Albers (2011) an keiner Stelle explizit angesprochen. Auch in einer früheren Veröffentlichung findet sich dazu kein Hinweis (siehe Tracy & Albers, 2006).

Trotz des offensichtlichen Bezugs zur Arbeit von Y. Miyake et al. (2004) verzichteten M. J. Albers (2011), sowie Tracy und Albers (2006) auf eine Inhibitionsaufgabe bzw. eine *rhythmische Klopfaufgabe* (d.h. bewusstes Pausemachen, siehe Park & Brünken, 2015) und instruierten stattdessen ihre Systemnutzer einen gleichmäßigen Rhythmus ohne Pausen (d.h. *regelmäßige Klopfaufgabe*) mit einer Geschwindigkeit von einem Schlag pro Sekunde mit ihrer nicht dominanten Hand zu klopfen. M. J. Albers (2011) argumentierte hier, dass mithilfe einer langsamen Schlagfolge, Ermüdungserscheinungen beim Anwender entgegengewirkt werden kann. Systemnutzern wurde beim regelmäßigen Klopfen freigestellt, ob sie mit der ganzen Hand oder nur einem Finger klopfen wollten. Sollte es bei der Bearbeitung der Webseite aufgrund einer Texteingabe erforderlich sein, beide Hände zu benötigen, kann in einer solchen Situation vom Nutzer auch mit dem Fuß geklopft werden. Während der eigentlichen Aufgabenbearbeitung wurde dann das Klopfgeräusch mithilfe einer normalen Bildschirmaufzeichnungssoftware akustisch aufgezeichnet. Im Anschluss wurde es einem Experten überlassen, die Videoaufzeichnung zu analysieren und diejenigen Punkte zu markieren, bei denen er glaubte, dass das Tempo des aufgezeichneten Klopfens langsamer oder schneller wurde, was laut M. J. Albers (2011) ein Anzeichen für eine Erhöhung der mentalen Beanspruchung des Nutzers darstellt. Auf Basis der Markierungen im Video konnte der Experte im Anschluss ein retrospektives Interview mit dem Systemnutzer durchführen, um an den markierten Videostellen Kausalitäten für eine erhöhte mentale Beanspruchung zu erkennen und die damit verbundenen Gestaltungselemente (z.B. Eingabefeld) auf Basis eines Think-Aloud-Protokolls zu identifizieren. Im Vergleich zu einem retrospektiven Think-Aloud-Protokoll müssen durch dieses Vorgehen nicht die ganzen Videoaufzeichnungen analysiert werden, was der zeitlichen Anwendungseffizienz der Methode zugutekommt. Die Veränderung in der Geschwindigkeit bei der regelmäßigen Klopfaufgabe stellt zusammenfassend daher die Entscheidungsgrundlage dar, auf deren Basis man entscheiden kann, ob sich hinter einer Rhythmusabweichung ein Nutzungsproblem verbirgt oder nicht. Es handelt sich dabei jedoch um ein subjektives Kriterium, das gänzlich der Einschätzung des Evaluators überlassen ist.

Im Vergleich zu den vorgestellten Arbeiten aus der Lernforschung im Zuge der Beschreibung der summativen Evaluation mit Rhythmusabweichungen (siehe Abschnitt 4.1) führten M. J. Albers (2011) sowie Tracy und Albers (2006) keine Meta-Evaluation ihrer formativen Methode durch. Sie untermauerten die Inhaltsvalidität von regelmäßigen Klopfaufgaben als geeignete Zweitaufgaben überwiegend theoretisch und führten stattdessen einen Pilottest mit 18 Probanden für eine interne Website eines Unternehmens durch. Unabhängig davon, dass für die Sicherstellung der wissenschaftlichen Güte der Methode von den Autoren nur die Inhaltsvalidität thematisiert wurde, kann ihre Methode auch an sich kritisiert werden. In keiner der beiden genannten Veröffentlichungen beschrieben die Wissenschaftler objektive Schwellenwerte für den Evaluator, anhand derer er möglichst objektiv entscheiden kann, ob das Tempo des Klopfens wirklich zu langsam oder zu schnell

ist. Die Entscheidung, im Video eine Markierung für das spätere retrospektive Interview zu setzen, ist somit vollkommen dem Gefühl des Evaluators überlassen. Dieser Aspekt der Methode ist sehr fragwürdig, da sogar professionelle Musiker Schwierigkeiten haben, das Tempo einer Schlagfolge valide einzuschätzen (Kuhn, 1974; Madsen, 1979).

Leider liegen in beiden Arbeiten (M. J. Albers, 2011; Tracy & Albers, 2006) darüber hinaus keine Angaben bezüglich der Übereinstimmung der Evaluatoren bezüglich der identifizierten Nutzungsprobleme vor, da der angesprochene Pilottest lediglich von einem einzelnen Evaluator durchgeführt wurde. Zusätzlich kann generell bereits die Verwendung einer regelmäßigen Klopfaufgabe anstelle einer Inhibition erfordernden rhythmischen Klopfaufgabe zu Problemen führen. Sowohl außerhalb der HCI-Forschung (siehe Gawron, 2019) als auch direkt innerhalb der HCI-Forschung (z.B. M. J. Albers, 2011; Tracy & Albers, 2006) kommen Klopfaufgaben überwiegend in Form von regelmäßigen Klopfaufgaben (d.h. keine Inhibition durch bewusste Pausen) zum Einsatz. Diese Art von Klopfaufgaben sind daher leicht zu automatisieren (siehe Park & Brünken, 2015). Dieses Phänomen der Automatisierung veranlasste womöglich einige Autoren in verschiedenen Forschungsbereichen zu der Annahme, dass eine zusätzliche motorische Aufgabe wie regelmäßiges Klopfen eine negative Auswirkung auf die kognitive Verarbeitung und die Performance in motorischen oder visuellen Hauptaufgaben hat (R. G. Brown & Marsden, 1991; Emerson & Miyake, 2003; Hegarty, Shah, & Miyake, 2000; Park & Brünken, 2015).

So zeigte sich beispielsweise in einer Studie von Hegarty et al. (2000) recht anschaulich, dass sich die Performanz der Teilnehmer bei einer etablierten Papierfaltaufgabe (siehe Ekstrom, Dermen, & Harman, 1976) beim parallelen regelmäßigen Klopfen sogar verbesserte, wohingegen sie bei anderen Zweitaufgaben (z.B. Generierung von Zufallszahlen) erwartungsgemäß abnahm. In einer anderen Studie konnten R. G. Brown und Marsden (1991) ebenfalls einen positiven Zusammenhang zwischen der Performance in motorischen Hauptaufgaben (z.B. Drücken eines Knopfes im Rahmen einer Stroop-Aufgabe zur Anzeige der Farbe der dargebotenen Wörter) und der Performance bei einer regelmäßigen Klopfaufgabe identifizieren. Emerson und Miyake (2003) zeigten dieses Phänomen ebenfalls, indem sie eine Rechenaufgabe als Hauptaufgabe in drei unterschiedlichen Bedingungen (d.h. zwei unterschiedliche Zweitaufgaben und keine Zweitaufgabe als Kontrollbedingung) im Rahmen des Zweitaufgabenparadigmas untersuchten. Sie stellten dabei fest, dass bei parallelem regelmäßigem Klopfen im Vergleich zu paralleler mündlicher Äußerung (d.h. wiederholtes lautes Aussprechen von „a-b-c“) die Reaktionszeiten verbessert wurden. Die Reaktionszeiten bei parallelem regelmäßigem Klopfen unterschieden sich dabei nicht signifikant von der Kontrollbedingung, wohingegen die Reaktionszeiten bei der mündlichen Aufgabe signifikant langsamer als in der Kontrollbedingung waren. Dies weist auf eine Automatisierung der Klopfaufgabe im Vergleich zur mündlichen Aufgabe hin.

Laut Park und Brünken (2015) weisen solche Befunde im Großen und Ganzen darauf hin, dass die Verwendung von regelmäßigem Klopfen im Rahmen eines Zweitaufgabenparadigmas ohne Interferenz abläuft und sich dementsprechend leicht automatisieren lässt. Diese Automatisierung verhindert jedoch, dass sich Performanceänderungen in der Hauptaufgabe verlässlich in der Klopfaufgabe widerspiegeln können, weswegen die von der Hauptaufgabe verursachte mentale Beanspruchung nicht valide durch die Zweitaufgabe erfasst werden kann. Laut dem Modell multipler Ressourcen von Wickens (2008) minimieren sich die Ressourcenanforderungen der Klopfaufgabe durch die Automatisierung bei der

perzeptiv-kognitiven Informationsverarbeitung (d.h. Anforderung an das Arbeitsgedächtnis). Die Klopfaufgabe benötigt dadurch nur noch physische Ressourcen für die eigentliche Handlungsausführung. Daher verursacht die Klopfaufgabe weniger oder keine Interferenz bei der perzeptiv-kognitiven Informationsverarbeitung der Hauptaufgabe, weshalb sie die gesamte Beanspruchung des Arbeitsgedächtnisses auch nicht abbilden kann. Kommt es trotzdem zu einer Änderung im Klopfrhythmus (d.h. schneller oder langsamer) kann nicht mit Sicherheit ausgeschlossen werden, dass diese Veränderung nicht durch Ermüdung oder einer motorischen Interferenz (d.h. Scrollen und gleichzeitiges Klopfen mit der Hand) verursacht ist.

Beim methodischen Vorgehen von M. J. Albers (2011) und Tracy und Albers (2006) kommt noch erschwerend hinzu, dass die Forscher keine individuelle Baseline der Nutzer vor der eigentlichen Systemnutzung erhoben haben. Es kann somit leicht passieren, dass Probanden aufgrund mangelnder motorischer Koordinationsfähigkeiten kontinuierlich aus dem vorgegebenen Rhythmus kommen. Dies würde bei der retrospektiven Analyse zu vielen vermeintlichen Rhythmusabweichungen führen (d.h. Fehlalarmen: Evaluator macht unnötige Markierungen, die er durchsprechen muss), hinter denen sich überhaupt keine kritischen Ereignisse verbergen, die auf intuitive Benutzung beeinträchtigende Nutzungsprobleme hinweisen. Dies gefährdet nicht nur die wissenschaftliche Güte der Methode, sondern auch deren zeitliche Anwendungseffizienz. Im Gegensatz zur von M. J. Albers (2011) vorgeschlagenen regelmäßigen Klopfaufgabe erfordert die Rhythmusmethode von Park und Brünken (2015) eine stetige Überwachung des Prozesses und kann als rhythmische Klopfaufgabe dementsprechend nicht leicht automatisiert werden. Es werden Inhibitionsprozesse benötigt, um bewusst die Pausen des gelernten Rhythmus einhalten zu können. Wie bereits in Abschnitt 2.2 ausgiebig dargestellt, kann die mentale Beanspruchung über Inhibitionsprozesse gemessen werden, da diese dem kognitiven Abkoppeln zugrunde liegen. Situationen, in denen der Nutzer kognitives Abkoppeln einsetzen muss, deuten auf eine erhöhte mentale Beanspruchung und damit auf Probleme hin, die eine intuitive Benutzung des Systems beeinträchtigen. Aus diesem Grund bietet die von Park und Brünken (2015) vorgeschlagene Rhythmusmethode auch großes Potential für die formative Evaluation intuitiver Benutzung im HCI-Bereich.

4.2.2 Diskussion von Limitationen und Bewältigung durch IntuiBeat-F

Auf Basis der nicht vorhandenen Meta-Evaluation und der im letzten Teilabschnitt aufgezählten Einschränkungen kann der von M. J. Albers (2011) vorgeschlagenen formativen Evaluationsmethode keine wissenschaftliche Güte bezüglich intuitiver Benutzung attestiert werden. Das Erkennen einer Tempoveränderung dem Evaluator selbst zu überlassen, birgt, aufgrund der hohen musikalischen Anforderungen, ein hohes Risiko viele vermeintliche Rhythmusabweichungen zu identifizieren hinter denen sich keine Nutzungsprobleme verbergen, die die intuitive Benutzung mit dem System beeinträchtigen. Ebenso kann das Verwenden einer regelmäßigen Klopfaufgabe, die keine bewusste Inhibition (d.h. kognitives Abkoppeln) erfordert, leicht automatisiert werden, was sich ebenfalls in einer höheren Anzahl Fehlalarmen und somit negativ in der wissenschaftlichen Güte und der zeitlichen Anwendungseffizienz der Methode widerspiegelt.

Nichtsdestotrotz konnten die Arbeiten von M. J. Albers (2011), sowie Tracy und Albers (2006) skizzieren, wie sich Klopfaufgaben potentiell auch zur formativen Evaluation im HCI-Bereich einsetzen lassen. Der Rhythmusmethode konnte bereits bezüglich der summativen Evaluation eine hohe Reliabilität und Inhaltsvalidität zugesprochen werden und erste Indizien (siehe Korbach et al., 2018; Park & Brünken, 2015) sprechen auch für das Vorhandensein höherer Formen der Validität. Demzufolge ist ein vollständiger Nachweis der wissenschaftlichen Güte von Rhythmusabweichungen als potentielle Indikatoren für intuitive Benutzung wahrscheinlich. Daher sind Rhythmusabweichungen auch für die formative Evaluation intuitiver Benutzung prädestiniert, falls dort anstelle einer regelmäßigen Klopfaufgabe (siehe M. J. Albers, 2011) eine rhythmischen Inhibitionsaufgabe mit individueller Baseline (siehe Park & Brünken, 2015) zum Einsatz kommt. Um jedoch Rhythmusabweichungen für die formative Evaluation von User Interfaces (speziell 3D-CUI-Interaktionslösungen im Zuge des Anwenderprojekts 3D-GUIde) einsetzen zu können, müssen zur Sicherstellung ihrer wissenschaftlichen Güte und zeitlichen Anwendungseffizienz eine Reihe von Anforderungen erfüllt werden:

- Adaption der Rhythmusmethode als IntuiBeat-F im Bereich der Mensch-Maschine-Interaktion, um sie zur formativen, auf mentaler Effizienz basierenden Evaluation intuitiver Benutzung bei User Interfaces (speziell 3D-CUIs im Rahmen des Anwenderprojekts 3D-GUIde) zeitlich effizient einsetzen zu können.
- Bereitstellung einer von der Expertise des Evaluators unabhängigen Analyse von Rhythmusabweichungen (d.h. Algorithmus) zur Identifikation von Kausalitäten mithilfe von IntuiBeat-F, die mit hoher Sicherheit für die Erhöhung der mentalen Beanspruchung verantwortlich sind.
- Sicherstellung der wissenschaftlichen Güte von IntuiBeat-F als formative Evaluationsmethode für intuitive Benutzung bei User Interfaces (speziell 3D-CUIs) anhand der Hauptgütekriterien *Gründlichkeit*, *Gültigkeit* und *Zuverlässigkeit* auf Basis des in Abschnitt 3.2.3 identifizierten aktuellen Benchmarks zur formativen Evaluation intuitiver Benutzung, dem Nutzertest mit retrospektivem Think-Aloud-Protokoll.
- Sicherstellung der *zeitlichen Anwendungseffizienz* von IntuiBeat-F durch den Vergleich mit dem in Abschnitt 3.2.3 identifizierten aktuellen Benchmarks zur formativen Evaluation intuitiver Benutzung, dem Nutzertest mit retrospektivem Think-Aloud-Protokoll.

4.3 Konkretisierung der Forschungsfragen

Auf der Basis bisheriger Forschungsergebnisse zur wissenschaftlichen Güte der formativen und summativen Evaluation intuitiver Benutzung soll im Rahmen dieser Arbeit festgestellt werden, ob eine auf Rhythmusabweichungen basierende Evaluationsmethode namens IntuiBeat gleichermaßen als formative (als IntuiBeat-F) und als summative (als IntuiBeat-S) Evaluationsmethode geeignet ist und zur Evaluation von User Interfaces (speziell 3D-CUI-Interaktionslösungen im Rahmen des Anwenderprojekts 3D-GUIde) zeitlich effizient genutzt werden kann. Aufgrund der in diesem Zusammenhang noch ungeklärten Aspekte und Limitationen von vorherigen Studien ergeben sich eine Reihe von Forschungsfragen, die im Rahmen dieser Dissertation beantwortet werden möchten.

Forschungsfrage 1: Wie hoch ist die wissenschaftliche Güte von IntuiBeat-S als summative Evaluationsmethode für intuitive Benutzung, beurteilt anhand der formalen Hauptgütekriterien Objektivität, Reliabilität und Validität?

Wie in Teilabschnitt 3.2.5 beschrieben, muss zur Bestimmung der Höhe der wissenschaftlichen Güte von IntuiBeat-S als summative Evaluationsmethode diese bezüglich der drei summativen Hauptgütekriterien (d.h. Objektivität, Reliabilität, Validität) formal bewertet werden. Eine derartige Bewertung ist nötig, da die wissenschaftliche Güte von Rhythmusabweichungen als summativer Indikator intuitiver Benutzung außerhalb des HCI-Bereichs noch unvollständig nachgewiesen ist und die Höhe wissenschaftlicher Güte bislang noch nicht in der HCI (speziell bei 3D-CUIs) bestimmt werden konnte. Aktuelle Befunde zur wissenschaftlichen Güte des Maßes beschränken sich bisher lediglich auf die Evaluation mentaler Beanspruchung im Bereich des E-Learnings. Des Weiteren wurde in diesem Zusammenhang ebenfalls noch nicht untersucht, ob bei einer solchen Methode etwaige Konfundierungen durch die Musikalität der Nutzer bestehen. Eine Übertragbarkeit des Maßes auf den für das Anwenderprojekt 3D-GUIde relevanten Bereich der 3D-CUIs muss daher noch im Rahmen dieser Arbeit demonstriert werden, weswegen die von Park und Brünken (2015) vorgeschlagene Rhythmusmethode zunächst für die Anwendung im HCI-Bereich adaptiert werden muss. Diese Adaption IntuiBeat-S beinhaltet Anpassungen im Versuchsaufbau, sowie die Implementierung einer Software (d.h. IntuiBeat-Software), die den Evaluator bei der Durchführung, Auswertung und Interpretation der Testwerte unterstützt.

Forschungsfrage 2: Wie hoch ist die wissenschaftliche Güte von IntuiBeat-F als formative Evaluationsmethode für intuitive Benutzung, beurteilt anhand der formalen Hauptgütekriterien Gründlichkeit, Gültigkeit und Zuverlässigkeit?

Wie in Teilabschnitt 3.2.5 beschrieben, muss zur Bestimmung der Höhe der wissenschaftlichen Güte von IntuiBeat-F als formative Evaluationsmethode diese bezüglich der drei formativen Hauptgütekriterien (d.h. Gründlichkeit, Gültigkeit und Zuverlässigkeit) formal bewertet werden. So können Rhythmusabweichungen nicht nur als summativer Indikator intuitiver Benutzung fungieren, sondern erlauben im Zuge einer formativen Evaluation auch, Kausalitäten zu erkennen, die eine intuitive Benutzung beeinträchtigen. Eine Bewertung des Ausmaßes wissenschaftlicher Güte bezüglich der Hauptgütekriterien ist nötig, da die Höhe wissenschaftlicher Güte von Rhythmusabweichungen als formativer Indikator für reale Nutzungsprobleme (d.h. beeinträchtigen intuitive Benutzung) bislang noch nicht im HCI-Bereich (speziell bei 3D-CUIs) bestimmt werden konnte. Bisherige Arbeiten beschränkten sich im Usability-Bereich (d.h. nicht speziell intuitive Benutzung) anstelle von rhythmischen Klopfaufgaben nur auf regelmäßige Klopfaufgaben, die mit einer Reihe von methodischen Einschränkungen verbunden sind. Diese Einschränkungen wurden ausführlich im letzten Abschnitt besprochen. Eine Übertragbarkeit des Maßes auf den für das Anwenderprojekt 3D-GUIde relevanten Bereich der 3D-CUIs muss im Rahmen dieser Arbeit somit noch demonstriert werden. Daher sollen bei der Adaption der von M. J. Albers (2011) vorgeschlagenen formativen Evaluationsmethode als IntuiBeat-F anstelle der Regelmäßigkeit des Tempos, die auch zur summativen Evaluation genutzten Rhythmusabweichungen als Indikatoren zur Identifikation von intuitive Benutzung beeinträchtigender

Nutzungsproblemen genutzt werden. In diesen Zusammenhang ist es ebenfalls nötig, Anpassungen im Versuchsaufbau vorzunehmen (z.B. Erhebung einer individuellen Baseline), und eine Software zu implementieren, die den Evaluator bei der Durchführung, Auswertung und Interpretation der Testwerte unterstützt. Hierbei ist die Auswahl eines geeigneten Algorithmus zentral, der dem Evaluator wissenschaftlich valide anzeigt, an welchen Stellen der Systemnutzer aus dem Rhythmus gekommen ist, und ihm dabei hilft, die Ursache dieser Abweichung zu identifizieren.

Forschungsfrage 3: Wie hoch ist die zeitliche Anwendungseffizienz von IntuiBeat-S und IntuiBeat-F im Vergleich zu bereits vorhandenen Evaluationsmethoden für intuitive Benutzung?

Wie bereits in der Einleitung dieser Arbeit (siehe Kapitel 1) beschrieben, handelt es sich bei der Entwicklung der Interaktionspatterns in 3D-GUIde um einen zeitkritischen Prozess, weswegen eine große Anzahl von Interaktionslösungen, auf deren Basis dann Patterns festgeschrieben werden können, möglichst zeitlich effizient evaluiert werden müssen. Um IntuiBeat, sowohl als formative (als IntuiBeat-F) als auch als summative (als IntuiBeat-S) Methode, zeitliche Anwendungseffizienz und damit im Rahmen des Projekts praktische Güte zu attestieren, muss die Höhe der zeitlichen Anwendungseffizienz durch den Vergleich mit dem formativen (d.h. Nutzertest mit retrospektivem Think-Aloud-Protokoll) und dem summativen (d.h. CHAI-Methode) Benchmark sichergestellt werden.

4.4 Experimentelles Vorgehen

Die Beantwortung der ersten Forschungsfrage erfolgt anhand des ersten, zweiten und dritten Experiments. Hierzu wird das Ausmaß an intuitiver Benutzung bei der Verwendung verschiedener CUIs zur Erledigung typischer CUI-Aufgaben (z.B. Verschieben von Objekten) mit IntuiBeat-S und den fünf in Abschnitt 3.6 identifizierten Quasi-Außenkriterien QUESI, CHAI, SEA, Effektivität und NASA-RTLX beurteilt, wobei die CHAI-Methode laut aktueller Erkenntnislage den Benchmark bezüglich summativer Evaluation im Forschungsfeld bildet. Im Zuge des ersten Experiments erfolgt die Beschreibung der technischen Implementierung eines Tools zur einheitlichen Aufzeichnung und Auswertung der Rhythmusabweichungen (d.h. IntuiBeat-Software) und damit auch die Bereitstellung der konkreten Spezifikation der Adaption der Rhythmusmethode als IntuiBeat-S für den HCI-Bereich (d.h. Vorgehensweise beim Einsatz der IntuiBeat-Software zur summativen Evaluation).

Das Gütekriterium der Objektivität wird in den drei Experimenten nicht empirisch überprüft, da aufgrund der hohen Automatisierung und Standardisierung von IntuiBeat-S durch die für die Anwendung bereitgestellte IntuiBeat-Software, sowie aufgrund der Tatsache, dass bei jedem der drei Experimente andere Evaluatoren mit unterschiedlicher Vorerfahrung eingesetzt werden, das Vorhandensein des Gütekriteriums der Objektivität von IntuiBeat-S unter Berücksichtigung von Döring und Bortz (2016) bereits theoretisch begründet werden kann. Die Reliabilität von IntuiBeat-S wird im Rahmen der drei Experimente in Form der Testhalbierungsreliabilität als methodischer Zugang bewertet und

im Vergleich mit allgemeingültigen quantitativen Normen interpretiert (siehe Döring & Bortz, 2016). IntuiBeat-S kann auf Basis dieses Vergleichs Reliabilität attestiert werden, wenn die ermittelten Reliabilitätskoeffizienten als akzeptabel betrachtet werden können und damit deskriptiv über .80 liegen.

Dem Gütekriterium der Validität wird sich mithilfe der Konstruktvalidität im Sinne eines struktursuchenden Vorgehens (siehe Fiske, 1982; Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007) anhand signifikanter Zusammenhänge mit konstruktnahen Quasi-Außenkriterien (subjektiv: SEA, NASA-RTLX und QUESI; objektiv: CHAI-Methode, Effektivität) und nicht signifikanter Zusammenhänge mit konstruktfremden Quasi-Außenkriterien (objektiv: physische Effizienz und die zeitliche Effizienz bei der Handlungsdurchführung) operational angenähert, da die Konstruktvalidität die Synthese aus Kriteriums- und Inhaltsvalidität darstellt. Sie kompensiert daher auch die Einschränkungen dieser beiden Arten der Validität (Döring & Bortz, 2016). Die divergenten Quasi-Außenkriterien wurden auf Basis der Überlegungen von Teilabschnitt 2.1.1 ausgewählt, da man so intuitive Benutzung von ihrem übergeordneten Konzept Usability eindeutig abgrenzen kann (siehe Hurtienne, 2011). IntuiBeat-S kann Validität attestiert werden, wenn die Korrelationen mit den konstruktnahen Verfahren deskriptiv höher liegen, als bei den konstruktfernen Verfahren. Dabei müssen konvergente signifikante Korrelationskoeffizienten deskriptiv größer oder gleich .50 sein, um als hoch zu gelten (J. Cohen, 1988), und damit die konvergente Validität von IntuiBeat-S zu bestätigen (siehe Fiske, 1982; Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007). Divergente Validität kann IntuiBeat-S in Ergänzung nur attestiert werden, wenn keine signifikanten Zusammenhänge zwischen IntuiBeat-S und den konstruktfremden Quasi-Außenkriterien bestehen (siehe Fiske, 1982; Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007). Da die Validität von Rhythmusabweichungen auch von der Musikalität der Nutzer konfundiert sein kann, widmet sich das erste Experiment diesem Aspekt.

Die Beantwortung der zweiten Forschungsfrage erfolgt anhand des vierten, fünften, sechsten und siebten Experiments. Hierzu werden mithilfe von IntuiBeat-F und einem Nutzertest mit retrospektivem Think-Aloud-Protokoll (d.h. aktueller Benchmark zur formativen Evaluation im Forschungsfeld) intuitive Benutzung beeinträchtigende Nutzungsprobleme (d.h. reale Nutzungsprobleme) bei der Verwendung verschiedener Software-Anwendungen (überwiegend CUIs) zur Erledigung typischer Aufgaben (z.B. Rotation von Objekten bei CUI-Programmen) abgeleitet. Um auch zur formativen Evaluation intuitiver Benutzung eingesetzt werden und damit intuitive Benutzung beeinträchtigende Nutzungsprobleme mit hoher Sicherheit identifizieren zu können, erfolgt im Zuge des vierten Experiments eine Anpassung der IntuiBeat-Software durch das Hinzufügen eines Algorithmus. Dieser Algorithmus hebt kritische Ereignisse hervor, damit diese für den Evaluator im Rahmen eines retrospektiven Interviews als Anhaltspunkte fungieren können. Auf diese Weise erfolgt auch die Bereitstellung der konkreten Spezifikation der Adaption der Rhythmusmethode als IntuiBeat-F für den HCI-Bereich (d.h. Vorgehensweise beim Einsatz der IntuiBeat-Software zur formativen Evaluation).

Das Gütekriterium der Zuverlässigkeit wird in den letzten vier Experimenten nicht empirisch untersucht, da Hartson et al. (2001) empfehlen, sich der Zuverlässigkeit einer formativen Evaluationsmethode erst zu widmen, wenn deren Gründlichkeit und Gültigkeit bestätigt werden konnte. Dies hat den Hintergrund, dass eine hohe Zuverlässigkeit nur durch Standardisierung erreicht werden kann, was zu einer niedrigen Variabilität bei der

Durchführung der formativen Evaluation führt. Dies kann sich dann in der Anzahl der mit der Methode gefundenen Nutzungsprobleme, der damit erreichten Gründlichkeit und Gültigkeit widerspiegeln. Um IntuiBeat-F Gründlichkeit und Gültigkeit attestieren zu können, müssen beide Gütekriterien bei IntuiBeat-F im Vergleich zum als Außenkriterium fungierenden Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der formativen Evaluation signifikant höher ausgeprägt sein (d.h. Gründlichkeit: höherer Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der vorhandenen realen Nutzungsprobleme im getesteten System; Gültigkeit: höherer Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der von der jeweiligen Evaluationsmethode gefundenen Nutzungsprobleme).

Die Beantwortung der dritten Forschungsfrage erfolgt parallel zur Beantwortung der ersten und zweiten Forschungsfrage anhand des zweiten (d.h. zeitliche Anwendungseffizienz von IntuiBeat-S) und der letzten vier Experimente (d.h. zeitliche Anwendungseffizienz von IntuiBeat-F). Zur Sicherstellung der zeitlichen Anwendungseffizienz von IntuiBeat-S wird überprüft, ob IntuiBeat-S eine höhere zeitliche Anwendungseffizienz als die ebenfalls objektive CHAI-Methode besitzt, die im Bereich summativer Evaluation intuitiver Benutzung den Benchmark darstellt (siehe Abschnitt 3.6). Liegt die zeitliche Anwendungseffizienz von IntuiBeat-S signifikant höher als die der CHAI-Methode, kann IntuiBeat-S eine hohe zeitliche Anwendungseffizienz zugesprochen werden. Die Sicherstellung der zeitlichen Anwendungseffizienz von IntuiBeat-F wird überprüft, indem die zeitliche Anwendungseffizienz von IntuiBeat-F mit der zeitlichen Anwendungseffizienz des Nutzertests mit retrospektivem Think-Aloud-Protokoll verglichen wird, der im Bereich formativer Evaluation den Benchmark darstellt (siehe Abschnitt 3.5). Liegt die zeitliche Anwendungseffizienz von IntuiBeat-F signifikant höher als die des Nutzertests mit retrospektivem Think-Aloud-Protokoll, kann IntuiBeat-F eine hohe zeitliche Anwendungseffizienz als wichtiger Teilaspekt praktischer Güte zugesprochen werden.

5 Güte von IntuiBeat-S für die summative Evaluation intuitiver Benutzung

Im vorangegangenen Kapitel wurde vorgestellt, warum Rhythmusabweichungen zur summativen Evaluation intuitiver Benutzung bei User Interfaces im Bereich der Mensch-Maschine-Interaktion potentiell geeignet sind. Anhand von drei Experimenten soll nun in diesem Kapitel nachgewiesen werden, inwiefern sich die Rhythmusmethode im Bereich der Mensch-Maschine-Interaktion als IntuiBeat-S (d.h. Nutzung von Rhythmusabweichungen zur summativen Evaluation intuitiver Benutzung) adaptieren lässt, ob der Methode dabei wissenschaftliche Güte hinsichtlich der drei vorgestellten formalen Hauptgütekriterien Objektivität, Reliabilität und Validität attestiert und ihr dabei auch zeitliche Anwendungseffizienz als wichtiger Teilaspekt praktischer Güte aus Anwenderprojektsicht zugesprochen werden kann. Dieses Kapitel widmet sich dementsprechend der Beantwortung der ersten Forschungsfrage und des summativen Aspekts der dritten Forschungsfrage (d.h. betrifft IntuiBeat-S) dieser Arbeit:

Forschungsfrage 1 Wie hoch ist die wissenschaftliche Güte von IntuiBeat-S als summative Evaluationsmethode für intuitive Benutzung, beurteilt anhand der formalen Hauptgütekriterien Objektivität, Reliabilität und Validität?

Forschungsfrage 3 Wie hoch ist die zeitliche Anwendungseffizienz von IntuiBeat-S und IntuiBeat-F im Vergleich zu bereits vorhandenen Evaluationsmethoden für intuitive Benutzung?

Zur Beantwortung der ersten Forschungsfrage wurden in jedem der drei folgenden Experimente jeweils zwei unterschiedlich intuitiv benutzbare CUIs (d.h. unabhängige Variable *unterschiedlich intuitiv benutzbare Software*: weniger intuitiv benutzbare Software vs. stärker intuitiv benutzbare Software) miteinander bezüglich des *Ausmaßes an intuitiver Benutzung* (d.h. abhängige Variable) anhand von IntuiBeat-S (d.h. mittleren Baseline-Abweichung des kurzen Rhythmusintervalls, siehe Abschnitt 4.1) und weiteren als konvergente Quasi-Außenkriterien fungierenden Evaluationsmethoden verglichen (siehe Abschnitt 3.6). Darüber hinaus wurden die beiden CUIs zusätzlich noch mithilfe anderer als divergente Quasi-Außenkriterien fungierende Evaluationsmethoden verglichen, die lediglich die physische und zeitliche Effizienz bei der Systemnutzung betrachten, also Konstrukte die nicht mit intuitiver Benutzung in Verbindung gebracht werden (siehe Teilabschnitt 2.1.1).

Empirische Überprüfung der Reliabilität und Validität

Das erste der drei Experimente (siehe Abschnitt 5.1) verwendete für die Meta-Evaluation als Datengrundlage einen Nutzertest mit paralleler Rhythmusaufgabe, um IntuiBeat-S

überhaupt einsetzen zu können. Da es dadurch zur Konfundierung der anderen konvergenten Evaluationsmethoden durch eine etwaige Intrusion der parallelen Rhythmusaufgabe (z.B. Verzerrung der Beurteilung der Schnelligkeit im Rahmen der CHAI-Methode durch eine zusätzliche Zweitaufgabe) gekommen sein kann, berücksichtige das zweite der drei Experimente (siehe Abschnitt 5.2) bei diesem Vergleich zusätzlich die *Art des Nutzertests* (d.h. unabhängige Variable *Art des Nutzertests*: Nutzertest mit Rhythmusaufgabe vs. Nutzertest ohne Rhythmusaufgabe), der als Datengrundlage für die Anwendung der anderen als konvergente Quasi-Außenkriterien fungierenden Evaluationsmethoden zum Einsatz kam. IntuiBeat-S konnte auf diese Weise auch mit der CHAI-Methode, die als aktueller objektiver Benchmark im Bereich summativer Evaluation intuitiver Benutzung gilt (siehe Teilabschnitt 3.6.2) unter idealen Ausführungsbedingungen (d.h. Evaluation mithilfe der CHAI-Methode auf Basis eines gewöhnlichen Nutzertests ohne zusätzliche Zweitaufgabe anstelle eines Nutzertests mit paralleler Rhythmusaufgabe) bezüglich der *zeitlichen Anwendungseffizienz* verglichen werden. Anhand der drei Experimente konnte zusammenfassend neben der ersten Forschungsfrage auch der summative Aspekt der dritten Forschungsfrage beantwortet, sowie gleichzeitig die Konfundierung durch eine mögliche Intrusion der für IntuiBeat-S erforderlichen parallel ausgeführten Rhythmusaufgabe ausgeschlossen werden.

In den ersten beiden Experimenten (siehe Abschnitte 5.1 und 5.2) fungierten Studierende der Mensch-Computer-Systeme und der Medienkommunikation als Stichprobe. Aufgrund des Anforderungsprofils der Studiengänge und der im Laufe des Studiums vermittelten Inhalte kann bei dieser Stichprobe erwartet werden, dass lediglich eine geringe Vorerfahrung bei der Nutzung von CUIs vorhanden ist. Dementsprechend lässt sich mit hoher Wahrscheinlichkeit eine Erstnutzung in den Experimenten beobachten, bei der die intuitive Benutzung eines Systems besonders kritisch (Blackler, 2006, 2018; Blackler, Popovic, & Mahar, 2005) und aufgrund der hohen Komplexität von CUIs, die sie üblicherweise nur Expertennutzern zugänglich macht, besonders wünschenswert ist (Akers, 2010; Akers et al., 2012). Das dritte der drei Experimente (siehe Abschnitt 5.3) nutzte für diesen Vergleich abschließend eine heterogenere Stichprobe (d.h. CUI-Experten: Studierende anderer Studiengänge, in denen die Nutzung von CUIs stärker verbreitet und damit ein höheres Maß an Vorerfahrung vorhanden ist) und andere Aufzeichnungshardware (d.h. am Lehrstuhl für Psychologische Ergonomie entwickeltes quelloffenes Fußpedal „Taktschuh“ anstelle eines Pedals eines bestimmten Herstellers, so wie es bei den ersten beiden Experimenten zum Einsatz kam). Auf diese Weise soll sowohl die generelle Robustheit der Methode demonstriert als auch eine universelle Hardware für die Anwendung von IntuiBeat-S für künftige Evaluatoren bereitgestellt werden. Die Reliabilität wurde dementsprechend im ersten (gewöhnliches Pedal) und dritten Experiment (Pedal „Taktschuh“) ebenfalls untersucht.

Theoretische Begründung der Objektivität

Im Rahmen der drei Experimente wurden lediglich die wissenschaftlichen Gütekriterien *Reliabilität* und *Validität* empirisch überprüft. Das Gütekriterium der *Objektivität* wurde stattdessen aufgrund der hohen Automatisierung und Standardisierung von IntuiBeat-S durch die dafür bereitgestellte IntuiBeat-Software, sowie aufgrund der Tatsache, dass zusätzlich bei jedem der drei Experimente andere Evaluatoren mit unterschiedlicher Vor-

erfahrung eingesetzt wurden, lediglich theoretisch begründet. Beim ersten Experiment übernahmen unter Anleitung des Versuchsleiters (d.h. Verfasser dieser Dissertation) eine studentische Hilfskraft (B.Sc. Mensch-Computer-Systeme 6. Semester, männlich, Anfang Zwanzig) und der Versuchsleiter selbst (M.Sc. Human-Computer Interaction abgeschlossen, männlich, Anfang Dreißig) die Durchführung des Experiments. Die Durchführung des zweiten Experiments wurde unter Anleitung des Versuchsleiters von zwei Studierenden des Bachelorstudiengangs Mensch-Computer-Systeme im Rahmen eines Seminars geleistet (Evaluator 1: B.Sc. Mensch-Computer-Systeme 5. Semester, männlich, Anfang Zwanzig; Evaluator 2: B.Sc. Mensch-Computer-Systeme 5. Semester, männlich, Anfang Zwanzig). Im dritten Experiment führte unter Anleitung eine Studierende des Bachelorstudiengangs Informationsdesign (Evaluator 3: B.Sc. Informationsdesign 7. Semester, weiblich, Anfang Zwanzig) das Experiment durch. Nutzertests von mehreren unterschiedlichen Personen in Teams durchführen zu lassen und bei allen Nutzertests nicht immer denselben Evaluator einzusetzen, ist in der Praxis Usus (z.B. Molich & Dumas, 2008; Molich et al., 2004; Molich et al., 1999).

5.1 Experiment 1

Das erste Experiment betrachtet die erste Forschungsfrage und untersucht die wissenschaftliche Güte von IntuiBeat-S bezüglich der formalen Hauptgütekriterien *Objektivität*, *Reliabilität* und *Validität*. Hierzu wurde IntuiBeat-F mit weiteren, als konvergente, sowie divergente Quasi-Außenkriterien fungierenden, summativen Evaluationsmethoden bei der Nutzung eines weniger intuitiv benutzbaren (Fusion 360, Autodesk, 2017a) und eines stärker intuitiv benutzbaren (SketchUp Trimble Navigation Ltd., 2016) CUI verglichen, welche auf Basis einer Experteneinschätzung ausgewählt wurden.

5.1.1 Überprüfung der Objektivität von IntuiBeat-S

Laut Moosbrugger und Kelava (2007), sowie Döring und Bortz (2016) kann einer Methode Objektivität in Form von Durchführungs-, Auswertungs- und Interpretationsobjektivität attestiert werden, wenn diese nahezu vollständig standardisiert durchgeführt werden kann. Um subjektive Abweichungen einzelner Evaluatoren bei der Durchführung, Auswertung und Interpretation zu vermeiden, wurde bei der Adaption der Rhythmusmethode als IntuiBeat-S darauf geachtet, dass diese bei der Durchführung und Auswertung durch die bereitgestellte IntuiBeat-Software hoch automatisiert ist, sowie eine objektive Interpretation der Testwerte gestattet (d.h. niedrige intuitive Benutzung bei hohen Rhythmusabweichungen von der individuellen Baseline, hohe intuitive Benutzung bei niedrigen Rhythmusabweichungen von der individuellen Baseline), die kein besonderes Expertenwissen des Evaluators voraussetzt und individuelle Deutungen mit hoher Wahrscheinlichkeit ausschließt. Dementsprechend wurde die Objektivität von IntuiBeat-S als summative Evaluationsmethode als erfüllt angesehen.

5.1.2 Überprüfung der Reliabilität von IntuiBeat-S

Da trotz einer hohen Automatisierung eine Verzerrung des Testergebnisses durch Messfehler nicht ausgeschlossen werden kann (Moosbrugger & Kelava, 2007), wurde das Gütekriterium der Reliabilität im Rahmen des ersten Experiments empirisch untersucht. Als methodischen Zugang zur Beurteilung der Reliabilität wurde hierbei die Testhalbierungs-Reliabilität gewählt, da diese bereits in den beiden Vorgängerstudien zur Rhythmusmethode als Zugang genutzt wurde. Man konnte daher auf Vergleichswerte zurückgreifen (siehe Korbach et al., 2018; Park & Brünken, 2015). Aufgrund der Tatsache, dass man Lerneffekte bei der Untersuchung von CUIs nicht ausschließen konnte und insbesondere die im dritten Experiment eingesetzten Expertennutzer nicht langfristig zur Verfügung standen, eignete sich die Retest- und Paralleltest-Reliabilität nicht für die Bestimmung der Reliabilität. In solchen Fällen kann laut Moosbrugger und Kelava (2007) auf die Testhalbierungs-Reliabilität zurückgegriffen werden. Als konkrete Operationalisierung der Testhalbierungs-Reliabilität wurde im Rahmen des ersten Experiments Guttman Reliabilitätskoeffizienten gewählt, da sie im Gegensatz zu den Spearman-Brown-Reliabilitätskoeffizienten nicht unterstellen, dass die Varianz in beiden Testhälften gleich sein muss. Sie gelten deswegen als eine konservativere Schätzung der Testhalbierungs-Reliabilität (Lienert & Raatz, 1998). Die Interpretation der Reliabilitätskoeffizienten erfolgte anhand eines Vergleichs mit allgemeingültigen quantitativen Normen (siehe Döring & Bortz, 2016). Laut Döring und Bortz (2016) können Reliabilitätskoeffizienten über oder gleich .90 als hoch und Reliabilitätskoeffizienten über oder gleich .80 als akzeptabel angesehen werden. IntuiBeat-S wurde als summative Evaluationsmethode intuitiver Benutzung nur dann Reliabilität attestiert, wenn der Reliabilitätskoeffizient in allen Versuchsbedingungen über oder gleich .80 lag und damit im akzeptablen Bereich lag.

5.1.3 Überprüfung der Validität von IntuiBeat-S

Da es trotz einer fehlerfreien Messung möglich sein kann, dass IntuiBeat-S überhaupt nicht das Ausmaß an intuitiver Benutzung abbildet, wurde auch die Validität im Rahmen des ersten Experiments empirisch untersucht (siehe Moosbrugger & Kelava, 2007). Als methodischer Zugang zur Beurteilung der Validität wurde hierbei die Konstruktvalidität gewählt, da diese die Synthese aus Kriteriums- und Inhaltsvalidität darstellt. Sie kann daher auch die Einschränkungen dieser beiden Arten der Validität kompensieren (Döring & Bortz, 2016). Konnte mithilfe konvergenter Methoden ein Unterschied bei der intuitiven Benutzung zwischen den beiden unterschiedlich intuitiv benutzbaren Softwareanwendungen festgestellt werden, sollte sich mit IntuiBeat-S dieser Unterschied ebenfalls feststellen lassen, wenn es sich bei der Methode um ein konstruktvalides Instrument handelt. Für die explizite Prüfung der Konstruktvalidität von IntuiBeat-S wurde im Rahmen des ersten Experiments ein struktursuchendes Vorgehen (siehe Abschnitt 3.4) durch Überprüfung der konvergenten und divergenten Validität gewählt.

Das Ausmaß der konvergenten Validität wurde anhand von vier Korrelationskoeffizienten (d.h. konvergente Validitätskoeffizienten) zwischen dem Testergebnis von IntuiBeat-S und dem Testergebnis der weiteren intuitive Benutzung erfassenden Evaluationsmethoden

ermittelt, bei denen es sich um die in Abschnitt 3.6 identifizierten konvergenten Quasi-Außenkriterien handelte. Da sich intuitive Benutzung subjektiv anhand des Gefühls von Flüssigkeit und objektiv anhand der mentalen Effizienz bewerten lässt (siehe Abschnitt 2.2), sollten für eine möglichst vollständige Erfassung des Konstrukts intuitiver Benutzung sowohl subjektive als auch objektive Methoden für die konvergente Konstruktvalidierung genutzt werden. Als subjektive Methoden kamen die SEA-Skala, der NASA-RTLX und der QUESI zum Einsatz. Als objektive Methoden wurden die Messung der Effektivität und die CHAI-Methode eingesetzt. Um IntuiBeat-S konvergente Validität attestieren zu können, sollten hohe signifikante lineare Zusammenhänge zwischen den Rhythmusabweichungen und den Testwerten der anderen fünf Methoden bestehen (alle Zusammenhänge positiv > 0 außer Effektivität und QUESI, da umgekehrt kodiert). Laut J. Cohen (1988) sind Korrelationskoeffizienten über oder gleich $|.50|$ als hoch zu bezeichnen, weswegen die konvergente Validität von IntuiBeat-S als summative Evaluationsmethode intuitiver Benutzung auch nur in diesem Bereich als bestätigt angesehen wurde (siehe Fiske, 1982; Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007).

Das Ausmaß der divergenten Validität wurde anhand von zwei Korrelationskoeffizienten zum Nachweis der divergenten Validität (d.h. divergente Validitätskoeffizienten) zwischen dem Testergebnis von IntuiBeat-S und den anderen erhobenen nicht intuitive Benutzung erfassenden Evaluationsmethoden ermittelt, bei denen es sich um die in Abschnitt 2.1.1 identifizierten divergenten Quasi-Außenkriterien handelte. Als objektive Methoden kamen hierbei die Messung der physischen Effizienz bei der Systemnutzung und die Messung der zeitlichen Effizienz bei der motorischen Handlungsdurchführung zum Einsatz. Um IntuiBeat-S divergente Validität attestieren und diese als bestätigt ansehen zu können, sollten keine signifikanten linearen Zusammenhänge zwischen den Rhythmusabweichungen und den Testwerten der anderen beiden Methoden bestehen. Des Weiteren sollten die konvergenten Validitätskoeffizienten deskriptiv höher als die divergenten Validitätskoeffizienten liegen, um einer Methode letztendlich Konstruktvalidität im Sinne eines struktursuchenden Vorgehens attestieren zu können (siehe Fiske, 1982; Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007).

5.1.3.1 Hypothesen

Zusammenfassend lassen sich die folgenden Hypothesen zur Sicherstellung der Konstruktvalidität von IntuiBeat-S festhalten:

H1 (Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs) Bei der stärker intuitiv benutzbaren Software liegt das mit den verschiedenen konvergenten Evaluationsmethoden erfasste Ausmaß an intuitiver Benutzung höher, als bei der weniger intuitiv benutzbaren Software.

- **H1.A (IntuiBeat-S)** Bei der stärker intuitiv benutzbaren Software sind die Rhythmusabweichungen mit IntuiBeat-S geringer, als bei der weniger intuitiv benutzbaren Software.

- **H1.B (SEA)** Bei der stärker intuitiv benutzbaren Software sind die Ratings der SEA-Skala geringer, als bei der weniger intuitiv benutzbaren Software.

- **H1.C (NASA-RTLX)** Bei der stärker intuitiv benutzbaren Software sind die Gesamtratings des NASA-RTLX geringer, als bei der weniger intuitiv benutzbaren Software.

- **H1.D (QUESI)** Bei der stärker intuitiv benutzbaren Software sind die Gesamtratings des QUESI höher, als bei der weniger intuitiv benutzbaren Software.

- **H1.E (Effektivität)** Bei der stärker intuitiv benutzbaren Software sind die Anteile korrekt abgeschlossener Aufgaben an der Gesamtzahl abgeschlossener Aufgaben höher, als bei der weniger intuitiv benutzbaren Software.

- **H1.F (CHAI)** Bei der stärker intuitiv benutzbaren Software sind die Anteile aller intuitiven Mausklicks an der Gesamtzahl der getätigten Mausklicks bei der Systemnutzung höher, als bei der weniger intuitiv benutzbaren Software.

H2 (Überprüfung der konvergenten Validität) Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und dem mit den verschiedenen konvergenten Evaluationsmethoden erfassten Ausmaß an intuitiver Benutzung bestehen hohe lineare Zusammenhänge.

- **H2.A (SEA)** Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und den Ratings der SEA-Skala besteht ein positiver, hoher, linearer Zusammenhang.

- **H2.B (NASA-RTLX)** Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und den Gesamtratings des NASA-TLX besteht ein positiver, hoher, linearer Zusammenhang.

- **H2.C (QUESI)** Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und den Gesamtratings des QUESI besteht ein negativer, hoher, linearer Zusammenhang.

- **H2.D (Effektivität)** Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und dem Anteil korrekt abgeschlossener Aufgaben an der Gesamtzahl abgeschlossener Aufgaben besteht ein negativer, hoher, linearer Zusammenhang.

- **H2.E (CHAI)** Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und dem Anteil aller intuitiven Mausklicks an der Gesamtzahl der getätigten Mausklicks bei der Systemnutzung besteht ein positiver, hoher, linearer Zusammenhang.

H3 (Überprüfung der divergenten Validität) Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und den nicht das Ausmaß an intuitiver Benutzung erfassenden divergenten Evaluationsmethoden bestehen keine linearen Zusammenhänge.

- **H3.A (Physische Effizienz)** Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und der Anzahl der benötigten Klicks bei der Systemnutzung besteht kein linearer Zusammenhang.

- **H3.B (Zeitliche Effizienz)** Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und der Zeit der motorischen Systemnutzung besteht kein linearer Zusammenhang.

5.1.4 Methode

5.1.4.1 Teilnehmer

Für das erste Experiment wurden 46 Versuchspersonen über das Probandensystem des Instituts für Mensch-Computer-Medien an der Universität Würzburg rekrutiert. Da für die Meta-Evaluation von IntuiBeat-S nur vollständige Datensätze berücksichtigt werden konnten, mussten 14 Versuchspersonen von der Datenauswertung ausgeschlossen werden (unvollständige Videoaufzeichnungen: 6; unvollständige Fragebögen: 4; unvollständige Rhythmusaufzeichnungen: 4). Demzufolge konnten für die Meta-Evaluation von IntuiBeat-S 32 Versuchspersonen berücksichtigt werden, welche alle rechtsfüßig (d.h. der rechte Fuß stellte den dominanten Fuß dar und wurde für die Rhythmuseingabe genutzt) waren. Die Stichrobe setzte sich dabei aus 22 Frauen und 10 Männern zusammen. Das Durchschnittsalter betrug 21.94 Jahre ($SD = 3.69$). Es handelte sich bei allen Teilnehmern um Studierende der Julius-Maximilians-Universität Würzburg, wobei neun Personen Mensch-Computer-Systeme (28.10 %) und 23 Personen Medienkommunikation (71.90 %) im Bachelor studierten. Alle Versuchsteilnehmer wurden über das Probanden-System des Instituts Mensch-Computer-Medien über eine gesonderte Mail darauf hingewiesen, für den Versuch flache Sportschuhe zu tragen, um eine möglichst problemlose Rhythmuseingabe über das USB-Fußpedal zu ermöglichen. Für die Teilnahme an der Untersuchung bekam jede Versuchsperson eine Versuchspersonenstunde gutgeschrieben. Die mit einem TFQ (siehe Anhang B.1.2) gemessene Vorerfahrung der Versuchspersonen bezüglich der Nutzung von CUIs betrug im Durchschnitt .04 ($SD = .12$) bei einem Maximum von 6 und lag damit, wie bei der Stichprobe erwartet, im unteren Bereich. Alle Versuchspersonen besaßen demzufolge eine geringe Vorerfahrung mit CUIs. Alle Versuchspersonen gaben an, am Experiment freiwillig teilzunehmen.

5.1.4.2 Versuchsdesign

Für die Beantwortung der ersten Forschungsfrage wurde ein einfaktorielles experimentelles Between-Subjects Design genutzt. Die unabhängige Variable war die *unterschiedlich intuitiv benutzbare Software* mit den Ausprägungen: weniger intuitiv benutzbare Software vs. stärker intuitiv benutzbare Software. Die abhängige Variable stellte das *Ausmaß an intuitiver Benutzung* dar.

5.1.4.3 Versuchsmaterialien und Maße

Unterschiedlich intuitiv benutzbare Software

Zur Operationalisierung der unabhängigen Variable *unterschiedlich intuitiv benutzbare Software* wurden auf Basis einer qualitativen Experteneinschätzung ($N_{Experte}$= 5; Experten mit umfangreicher Vorerfahrung im Bereich Usability und in der CUI-Domäne; Nutzung der Checkliste *Evalint* mit der analytisch die intuitive Benutzung einer Software beurteilt werden kann (Mohs, Hurtienne, Scholz, & Rotting, 2006)) das 3D-CUI *Fusion*

360 von Autodesk (2017a) als wenig intuitiv benutzbare Software (d.h. erwartete hohe Diversität und Komplexität der Systemnutzung, da das Experten-Tool den Anspruch erhebt den vollständigen Gestaltungs-, Entwicklungs- und Fertigungsprozess abzubilden) und das 3D-CUI *SketchUp* von Trimble Navigation Ltd. (2016) als stärker intuitiv benutzbare Software (d.h. erwartete geringe Diversität und Komplexität, da das Laien-Tool lediglich den Gestaltungsprozess abbildet) als Untersuchungsgegenstände für die Meta-Evaluation von IntuiBeat-S ausgewählt (siehe Abbildung 5.1). Um diese beiden Ausprägungen miteinander vergleichen zu können, wurden von den gleichen Experten außerdem drei experimentelle Aufgaben gewählt, die allesamt in ein fiktives Szenarios eingebettet waren, in dem es um die Neugestaltung eines Raumes ging. Da mit jedem System die gleichen drei Aufgaben bearbeitet wurden, wurde sich entsprechend für ein Between-Subjects Design entschieden, um etwaige Übungseffekte vermeiden zu können.

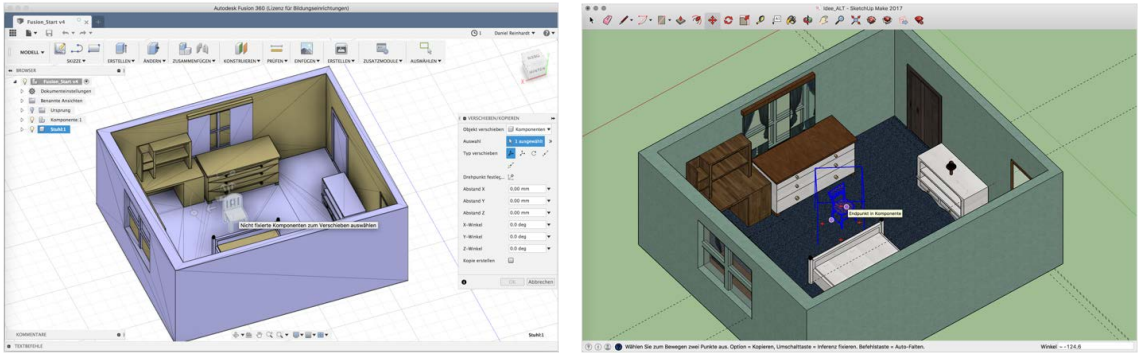


Abbildung 5.1. Die getesteten 3D-CUIs *Fusion 360* (links, Autodesk, 2017a) und *SketchUp* (rechts, Trimble Navigation Ltd., 2016), welche die beiden Ausprägungen (d.h. Fusion 360: weniger intuitiv benutzbare Software; SketchUp: stärker intuitiv benutzbare Software) der unabhängigen Variable *unterschiedlich intuitiv benutzbare Software* repräsentierten.

Bei den drei experimentellen Aufgaben (siehe Anhang B.1.1) handelte es sich um domänenübergreifende grundlegende Testaufgaben[1]. Sie wurden von den Versuchspersonen in einer randomisierten Reihenfolge bearbeitet. Die erste Aufgabe erforderte es, dass die Versuchspersonen die Position einer Lampe im Raum verändern mussten (erste Aufgabe). In der nächsten Aufgabe mussten die Versuchspersonen einen Stuhl um 180 Grad drehen (zweite Aufgabe). Die letzte Aufgabe erforderte es von den Versuchspersonen den Rahmen einer Tür zu messen (dritte Aufgabe).

Wie bereits in der Einleitung dieser Arbeit und im Teilabschnitt 3.1 beschrieben, brachte die im Rahmen des Projekts 3D-GUIde durchgeführte Nutzungskontextanalyse die Erkenntnis, dass die meisten Probleme bei der Handhabung von CUIs, bei den von LaViola et al. (2017) beschriebenen domänenübergreifenden, grundlegenden Nutzeraufgaben Selektion, Manipulation, Navigation, Zeicheneingaben und Systemsteuerung entstehen. Man benötigt daher hierfür künftig dringend intuitiv benutzbare 3D-CUI-Interaktionslösungen als Interaktionspatterns. Die für das erste und auch die weiteren Experimente als Stimuli gewählten Testaufgaben involvierten demzufolge gleich mehrere dieser grundlegenden Nutzeraufgaben, indem sich jede Aufgaben wiederum aus drei Teilaufgaben zusammensetzte.

[1] Diese Testaufgaben stellten mögliche Interaktionspatternkandidaten für das Projekt 3D-GUIde dar, über die im Rahmen der Untersuchung qualitative Daten gesammelt werden sollten.

Zu Beginn einer jeden Testaufgabe mussten sich die Versuchspersonen zunächst in der 3D-Szene zurecht finden und das für die Testaufgabe relevante Objekt ausfindig machen (d.h. Aufgaben der Navigation). Im Anschluss forderte jede Testaufgabe von den Versuchspersonen die entsprechende Funktion im Menü auszuwählen (d.h. Aufgaben der Systemsteuerung). Die abschließende Verwendung der gewählten Funktionalität implizierte dann eine indirekte Manipulation per Zeicheneingabe (z.B. Einstellen des Rotationsgrads durch Eingabe eines numerischen Werts in ein Textfeld) oder eine direkte Manipulation (z.B. Rotation des Objekts mithilfe eines Transformationsmanipulators, wodurch das Objekt direkt durch Ziehen des Objekts an Ecken und Kanten gedreht werden kann). Durch diese Struktur konnten die Testaufgaben zu evaluierende 3D-CUI-Interaktionslösungen möglichst realitätsgetreu abbilden und der Evaluator konnte im Rahmen eines retrospektiven qualitativen Interviews potentielle Nutzungsprobleme entdecken.

Summative Evaluation intuitiver Benutzung mit IntuiBeat-S

Die abhängige Variable *Ausmaß an intuitiver Benutzung* wurde im Rahmen des ersten Experiments sowohl mit objektiven konvergenten Evaluationsmethoden, die intuitive Benutzung anhand der mentalen Effizienz abbilden, als auch mit subjektiven konvergenten Evaluationsmethoden, die intuitive Benutzung anhand des wahrgenommenen Gefühls von Flüssigkeit bestimmen, operationalisiert. Als zentrale objektive Methode wurde IntuiBeat-S (d.h. AV_IntuiBeat-S) genutzt, da deren wissenschaftliche Güte im Rahmen des ersten Experiments wie auch dieser Arbeit demonstriert werden sollte. Wie bereits in Teilabschnitt 4.1 erwähnt, verbirgt sich hinter der Bezeichnung *IntuiBeat-S* eine im Vorfeld des ersten Experiments vorgenommene Adaption der Rhythmusmethode (siehe Korbach et al., 2018; Park & Brünken, 2015) für die summative Evaluation intuitiver Benutzung im HCI-Bereich. Wie die ursprüngliche Rhythmusmethode operationalisiert auch diese Adaption die mentale Beanspruchung bzw. mentale Effizienz der kognitiven Informationsverarbeitung während der Systemnutzung objektiv anhand der mittleren Baseline-Abweichung des kurzen Rhythmusintervalls. Im Folgenden soll nun die Vorgehensweise von IntuiBeat-S zur summativen Evaluation intuitiver Benutzung anhand der mentalen Effizienz und des Einsatzes der dafür erforderlichen Testumgebung vor dem Hintergrund des ersten Experiments beschrieben werden. Diese Beschreibungen lassen sich jedoch generalisieren und fungieren damit als eine generelle Methodenbeschreibung von IntuiBeat-S.

Hard- und Software von IntuiBeat-S

Die beiden vorgestellten Arbeiten zur Rhythmusmethode nutzten für die Aufzeichnung des Rhythmus entweder die Audioaufnahme-Software *Audacity* von Mazzoni (2014) und ein von den Autoren nicht weiter spezifiziertes Fußpedal (siehe Park & Brünken, 2015) oder die Experimentalsteuerungssoftware *E-Prime* zusammen mit einer Serial-Response-Box und einem dazugehörigen Fußpedal (siehe Korbach et al., 2018). Im Gegensatz zu diesen beiden Arbeiten wurde für die Evaluation im HCI-Bereich ein neues universelles Tool (d.h. IntuiBeat-Software) für die Aufzeichnung und Auswertung von Rhythmusabweichungen entwickelt. Da im Projekt 3D-GUIde die 3D-CUI-Interaktionslösungen von verschiedenen Evaluatoren an unterschiedlichen Standorten getestet werden sollen, war

dementsprechend ein im Vergleich zu den Originalarbeiten einfacheres Setup für die Anwendung der Rhythmusmethode wünschenswert. Dieses sollte am besten mit (1) besser auf einander abgestimmter Standardkomponenten (d.h. Rhythmusaufzeichnung mit herkömmlichen USB-Fußpedal statt Spezialpedal) auskommen, (2) eine hohe Automatisierung bei der Datenaufzeichnung und -auswertung zulassen (d.h. Sicherstellung der Objektivität durch unterstützende Software), (3) mit jedem Betriebssystem nutzbar sein und (4) generell Plug und Play erlauben.

Das von Park und Brünken (2015) genutzte Setup mit Audacity und dem nicht weiter spezifizierten Fußpedal scheint diese Anforderungen jedoch nicht vollständig zu erfüllen, da davon auszugehen ist, dass eine hohe Automatisierung mit diesem Setup schwierig ist. Man kann darüber hinaus auf Basis der Veröffentlichung von Park und Brünken (2015) nicht genau einschätzen, wie das Fußpedal überhaupt für eine reibungslose Kommunikation mit Audacity gestaltet sein muss. Das E-Prime-Setup, das im zweiten Experiment zur Rhythmusmethode von Korbach et al. (2018) genutzt wurde, lässt sich vorwiegend aufgrund der damit verbundenen Kosten von mehreren Hundert Euro und der Beschränkung auf Windows als Betriebssystem nicht so einfach an verschiedenen Orten von mehreren Evaluatoren einsetzen. Eine uneingeschränkte Eignung für eine Anwendung im Projekt 3D-GUIde ist daher fragwürdig.

Deswegen wurde für die Adaption der Rhythmusmethode als IntuiBeat-S eine entsprechende Testumgebung bereitgestellt, die sich aus einer in Java (Version 1.8) geschriebenen Software zur Aufzeichnung und Auswertung von Rhythmusabweichungen (d.h. IntuiBeat-Software) und einem handelsüblichen USB-Fußpedal (d.h. IntuiBeat-Hardware, siehe Abbildung 5.2), das den von der Versuchspersonen geklopften Rhythmus aufzeichnen kann, zusammensetzt. Darüber hinaus musste für die Adaption im HCI-Bereich das ursprüngliche Vorgehen der Rhythmusmethode bei der Aufzeichnung von Rhythmusabweichungen an einigen Stellen modifiziert werden (d.h. IntuiBeat-S), um auf diese Weise ein einfacheres Setup bereitstellen zu können, das die oben beschriebenen Anforderungen erfüllen kann.

Abbildung 5.2. Das im Rahmen des ersten und zweiten Experiments genutzte USB-Fußpedal (Typ FS1-P) der Firma Sodial.

Die IntuiBeat-Software wurde dem Evaluator auf seinem Rechner (siehe Analyse-PC in Absatz „Apparatur“) als JAR-Datei bereitgestellt. Für die Aufzeichnung des Rhythmus

wurde ein USB-Fußpedal (Länge circa 99 Millimeter, Breite circa 63 Millimeter, Höhe circa 40 Millimeter) der Firma Sodial (Typ FS1-P) verwendet (siehe Abbildung 5.2). Dem Fußpedal musste beim erstmaligen Anschließen über die beiliegende Treibersoftware ein Hotkey zugewiesen werden, da ein solches USB-Fußpedal nichts anderes als eine externe Tastatur mit lediglich einer einzelnen Taste ist, die durch das Drücken des Pedals angestoßen wird. Im Falle des ersten Experiments wurde dem Pedal der Hotkey „0“ zugewiesen, um die Eingabe des Pedals durch die in Java geschriebene Software erkennen zu können. Jedes Drücken des Fußpedals wurde von der Software als Schlag bzw. Eingabe interpretiert und entweder dem kurzen oder dem langen Rhythmusintervall zugeordnet. Zwei aufeinander folgenden Rhythmusintervalle bildeten eine sogenannte Rhythmuseinheit (Korbach et al., 2018; Park & Brünken, 2015). Um das Ausmaß an intuitiver Benutzung anhand von Rhythmusabweichungen überhaupt summativ evaluieren zu können, wurde vom Evaluator zunächst die individuelle Baseline der Versuchspersonen in Anlehnung an die Originalarbeiten zur Rhythmusmethode (siehe Korbach et al., 2018; Park & Brünken, 2015) erhoben. Dazu schloss er zunächst das USB-Pedal an seinem Rechner an (siehe Analyse-PC in „Apparatur“) und öffnete dort die IntuiBeat-Software.

Funktionsweise des Baseline-Modus der IntuiBeat-Software und Beschreibung des Programmablaufs

Im Startdialog wechselte er dann durch Klicken auf die Schaltfläche „Ja“ in den *Baseline-Modus* (siehe Abbildung 5.3, Schritt 1). In diesem Modus vergab er zunächst eine eindeutige ID für die Baseline-Messung (siehe Abbildung 5.3, Schritt 2). An dieser Stelle hat es sich bewährt als ID immer das Versuchspersonenkürzel zu verwenden. Nach der Eingabe der ID setzte die Software intern die Dauer des kurzen Rhythmusintervalls auf 500 und das lange Rhythmusintervall auf 1500 Millisekunden, was einer gesamten Rhythmuseinheit von 2000 Millisekunden und somit dem von Park und Brünken (2015) ursprünglich vorgeschlagenen Rhythmus im Viervierteltakt entspricht.

Im Anschluss öffnete sich das Hauptfenster des Baseline-Modus innerhalb der IntuiBeat-Software, womit der Evaluator nun die eigentliche Baseline-Messung der Versuchspersonen durchführen konnte (siehe Abbildung 5.3, Schritt 3).

1. Im Baseline-Modus wurden die Versuchspersonen vom Evaluator zunächst standardisiert mündlich instruiert, sich den Rhythmus ohne eigene Aktivität (d.h. kein selbstständiges Klopfen auf das Fußpedal) als Audioaufzeichnung vorspielen zu lassen. Das Hauptfenster des Baseline-Modus bot für diesen Zweck dem Evaluator die Möglichkeit, den Rhythmus im Viervierteltakt den Versuchspersonen für 10 Sekunden durch Klicken auf die Schaltfläche „10 s Ton abspielen“ als Audioaufzeichnung vorzuspielen (siehe Abbildung 5.3, Schritt 3).
2. Damit die Versuchspersonen den Rhythmus auch so verinnerlichen konnten, dass die spätere Inhibition des Rhythmus aufgrund der damit verbundenen Erhöhung der mentalen Beanspruchung im Sinne eines Zweitaufgaben-Paradigmas während der Systemnutzung feststellbar ist (siehe Abschnitt 2.2), wurden die Versuchspersonen vom Evaluator als nächstes standardisiert mündlich instruiert, den Rhythmus im folgenden Übungsmodus synchron mit der folgenden Audioaufzeichnung mit dem

USB-Pedal mitzuklopfen. Das Hauptfenster des Baseline-Modus bot für diesen Zweck dem Evaluator die Möglichkeit, den Rhythmus im Viervierteltakt den Versuchspersonen durch Klick auf die Schaltfläche „30 s Ton abspielen“ für 30 Sekunden als Audioaufzeichnung zur Unterstützung vorzuspielen (siehe Abbildung 5.3, Schritt 3). Während die Versuchspersonen versuchten den Rhythmus synchron zur Audioaufzeichnung mitzuklopfen, registrierte die IntuiBeat-Software in diesem Übungsmodus jede Eingabe über das Fußpedal und gab den Versuchspersonen akustisches Feedback in Form eines kurzen Klopf-Tons über den Rechner des Evaluators. Auf diese Weise wurde den Versuchspersonen eine erkannte Eingabe signalisiert.

3. Nachdem die Versuchspersonen den Rhythmus auf die beschriebene Weise üben konnten, konnte die Baseline der Versuchspersonen (d.h. Rhythmusaufzeichnung ohne zusätzliche mentale Belastung durch die eigentliche Systemnutzung) vom Evaluator mithilfe der IntuiBeat-Software erhoben werden. Versuchspersonen wurden zu diesem Zweck zunächst vom Evaluator standardisiert mündlich instruiert, den gelernten Rhythmus nun selbstständig ohne zusätzliche Audioaufzeichnung wie beim Übungsmodus, jedoch ebenfalls mit akustischen Feedback in Form eines kurzen Klopf-Tons für 60 Sekunden selbstständig zu klopfen. Dieses selbstständige Klopfen stellte damit die Baseline der Versuchsperson dar. Das Hauptfenster des Baseline-Modus bot dem Evaluator für diesen Zweck die Möglichkeit, die Rhythmusaufzeichnung zu starten, zu beenden und zu speichern (siehe Abbildung 5.3, Schritt 3).

Nachdem die Versuchspersonen dem Evaluator mündlich signalisiert hatten, dass sie seine Instruktion verstanden hatten und mit der Erhebung ihrer Baseline beginnen möchten, startete der Evaluator mithilfe der IntuiBeat-Software über die Schaltfläche „Aufnahme starten“ die Baseline-Aufnahme. Nach Betätigung der Schaltfläche signalisierte die Software dem Evaluator die nun laufende Aufnahme mit den Worten „Aufnahme läuft“ (siehe Abbildung 5.3, Schritt 4).

Wie in Teilabschnitt 4.1 ausführlich unter Berücksichtigung der bisherigen Originalarbeiten zur Rhythmusmethode (siehe Korbach et al., 2018; Park & Brünken, 2015) begründet, müssten bei einer Baseline-Aufzeichnung alle Eingaben mit einem USB-Fußpedal, die weniger als 250 Millisekunden auseinander liegen, für die Auswertung direkt verworfen werden. Die IntuiBeat-Software machte dies automatisch. Ebenso verworf die Software Eingaben, die mehr als 2000 Millisekunden auseinander lagen, da eine Rhythmuseinheit des von Park und Brünken (2015) vorgeschlagenen Rhythmus nur insgesamt 2000 Millisekunden umfassen darf. Eingaben, die auf diese Weise nicht verworfen wurden und größer als 1000 Millisekunden waren, wurden von der Software dem langen Rhythmusintervall zugeordnet. Die restlichen Eingaben wurden dem kurzen Rhythmusintervall zugeordnet. Da eine Rhythmuseinheit immer aus einem kurzen und einem langen Rhythmusintervall bestehen muss, wurden Einheiten, die durch zwei Intervalle des gleichen Typs gebildet wurden (d.h. zwei lange oder zwei kurze Rhythmusintervalle) von der Software entsprechend verworfen (siehe Abschnitt 4.1). Der Evaluator wurde während der Baseline-Aufzeichnung über die Gültigkeit einer Rhythmuseinheit informiert, indem sich bei einer ungültigen Einheit die Worte „Aufnahme läuft“ rot (siehe Abbildung 5.3, Schritt 4a) färbten. Wie bereits erwähnt, bekamen die Versuchspersonen während der Baseline-Messung so wie im Übungsmodus durch ein kurzes Klopf-Tons nach jeder Eingabe Feedback von der Software, dass diese Eingabe von ihr erfolgreich registriert wurde.

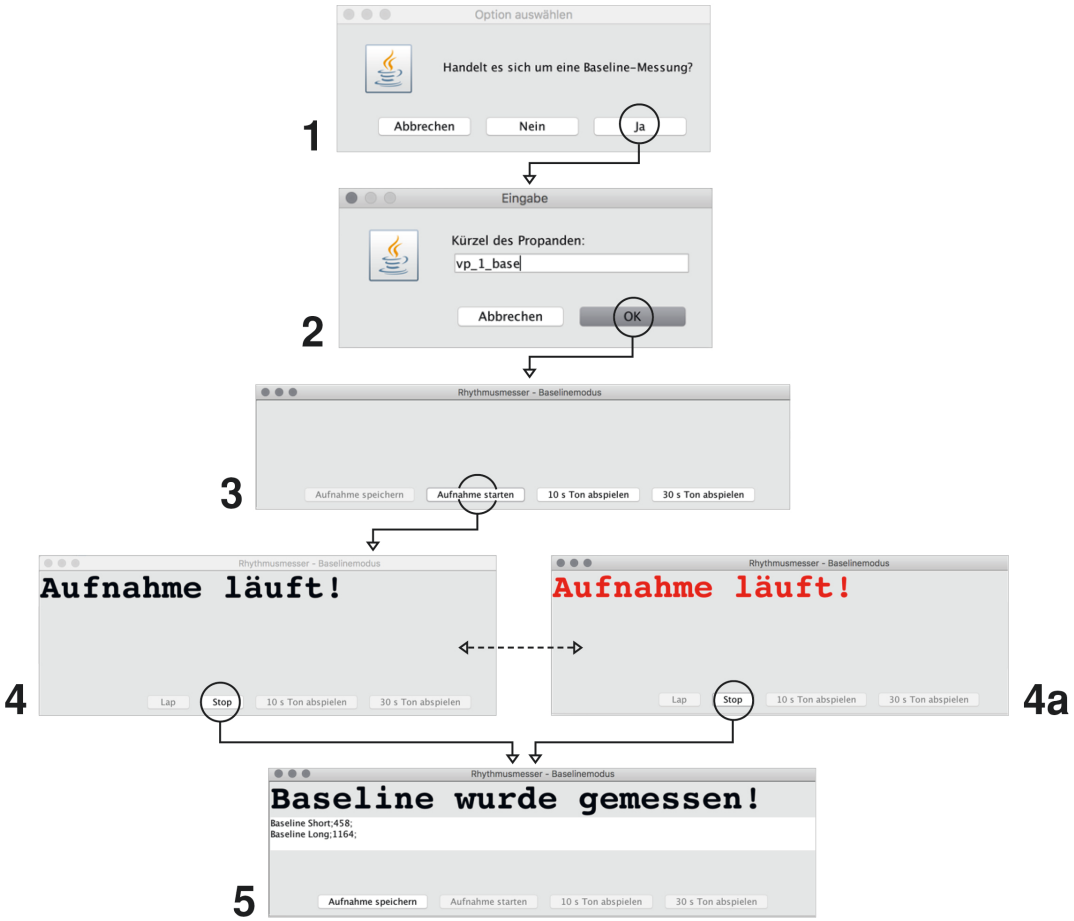


Abbildung 5.3. Flussdiagramm des Baselinemodus der IntuiBeat-Software.

Nach 60 Sekunden stoppte die Baseline-Aufzeichnung automatisch und die Software zeigte dem Evaluator die tatsächliche Länge des kurzen und langen Rhythmusintervalls als die erhobene individuelle Baseline der Versuchsperson an (siehe Abbildung 5.3, Schritt 5). An dieser Stelle ist anzumerken, dass der Evaluator auch die Aufnahme manuell durch Klicken auf die Schaltfläche „Aufnahme stoppen“ hätte beenden können. Abschließend speicherte der Evaluator die Baseline-Aufzeichnung durch Klicken auf die Schaltfläche „Aufnahme speichern“ (siehe Abbildung 5.3, Schritt 5). Im Anschluss wurden von der Software drei verschiedene Logs als CSV-Dateien generiert (d.h. Typ „[ID]_raw.csv“, Typ „[ID]_clean.csv“, Typ „[ID]_base.csv“), die im selben Ordner auf dem Rechner des Evaluators (d.h. Analyse-PC, siehe Absatz „Apparatur“) abgelegt wurden, auf dem sich auch die JAR-Datei der IntuiBeat-Software befand. Die Software beendete sich dann automatisch selbst. Es wurde sich bei der Gestaltung der Software entschieden dem Evaluator die Aufzeichnung explizit speichern zu lassen, da es bei der Rhythmusaufzeichnung zu Fehlern gekommen sein kann (z.B. Fehlfunktion des USB-Fußpedals), weswegen die Möglichkeit bestehen könnte, dass die Aufnahme verworfen werden muss. Eine Aufnahme wurde dadurch verworfen, dass der Evaluator diese nicht explizit speicherte, sondern stattdessen einfach die Software beende-

te. Um zu vermeiden, dass der Evaluator die Anwendung unbeabsichtigt beendet, öffnete sich vor dem eigentlichen Schließen noch ein Bestätigungsdialog.

Eine Raw-Datei enthielt dabei die Zeitstempel der Eingaben mit dem Fußpedal in Millisekunden. Wie bereits erwähnt werden von der Software zwei aufeinanderfolgende Intervalle (d.h. vier Eingaben: die ersten beiden Eingaben definierten das erste und die beiden letzten Eingaben das zweite Intervall) zu einer Rhythmuseinheit zusammengefasst, was einer Zeile in einer Raw-Datei entspricht. Eine Clean-Datei enthielt nur die gültigen Rhythmuseinheiten (d.h. Rhythmuseinheiten, die den vorgegeben Rhythmus im Viervierteltakt unter den oben genannten Ausschlusskriterien überhaupt reproduzieren konnten) bestehend aus einem kurzen und einem langen Rhythmusintervall, die für die Berechnung der Baseline (d.h. nur kurzes Rhythmusintervall) genutzt wurden. Eine Baseline-Datei enthielt die Baseline für das kurze (d.h. Baseline *kurz*) und lange (d.h. Baseline *lang*) Rhythmusintervall, welche die Mittelwerte über die entsprechenden Intervalle der Clean-Datei darstellen. Wie bereits in Teilabschnitt 4.1 erwähnt, verwendet IntuiBeat-S für die Beurteilung des Ausmaßes an intuitiver Benutzung nur das kurze Intervall als Baseline für die Berechnung der Rhythmusabweichungen zur summativen Evaluation intuitiver Benutzung. Bei der Gestaltung der Software hat man sich jedoch für die Aufzeichnung von beiden Intervallen entschieden, um diese Daten bei etwaigen explorativen Datenanalysen zur Verfügung zu haben. Diese Analysen sind jedoch nicht Gegenstand dieser Arbeit. Im Anschluss an die Messung der Baseline, konnte der Evaluator für die Durchführung jeder experimentellen Aufgabe das Ausmaß intuitiver Benutzung anhand der mentalen Beanspruchung der Versuchspersonen auf Basis ihrer erhobenen individuellen Baselines evaluieren.

Funktionsweise des Experimental-Modus der IntuiBeat-Software und Beschreibung des Programmablaufs

Dazu öffnete der Evaluator bei der Bewertung jeder Aufgabe die IntuiBeat-Software im *Experimental-Modus*, den er nach dem erneuten Starten der Software auf seinem Rechner durch Klicken auf die Schaltfläche „Nein“ im Startdialog erreichte (siehe Abbildung 5.4, Schritt 1). Im Folgenden vergab der Evaluator, wie bereits bei der Baseline-Messung zuvor, eine eindeutige ID für die Aufzeichnung (siehe Abbildung 5.4, Schritt 2). An dieser Stelle hat es sich bewährt für eine ID das Versuchspersonenkürzel mit der Aufgaben-ID (z.B. Positionierung für die Positionierungsaufgabe) zu kombinieren. Daraufhin gab der Evaluator die Dauer des kurzen und langen Rhythmusintervalls in zwei Folgedialogen an (siehe Abbildung 5.4, Schritt 3 und 4). Diese Parameter entnahm der Evaluator der zuvor erhobenen Baseline-Datei. Wie bereits erwähnt, benötigt IntuiBeat-S für die summative Beurteilung des Ausmaßes an intuitiver Benutzung nur das kurze Rhythmusintervall, weswegen man sich bei der Software die Eingabe der langen Rhythmus-Baseline hätte sparen können. Um jedoch die Daten für eine potentielle explorative Datenanalyse zur Verfügung zu haben, wurde bei der Gestaltung der Software entschieden, dass diese Baseline zur Berechnung der Rhythmusabweichungen des langen Rhythmusintervalls von der Software zusätzlich erfragt wird.

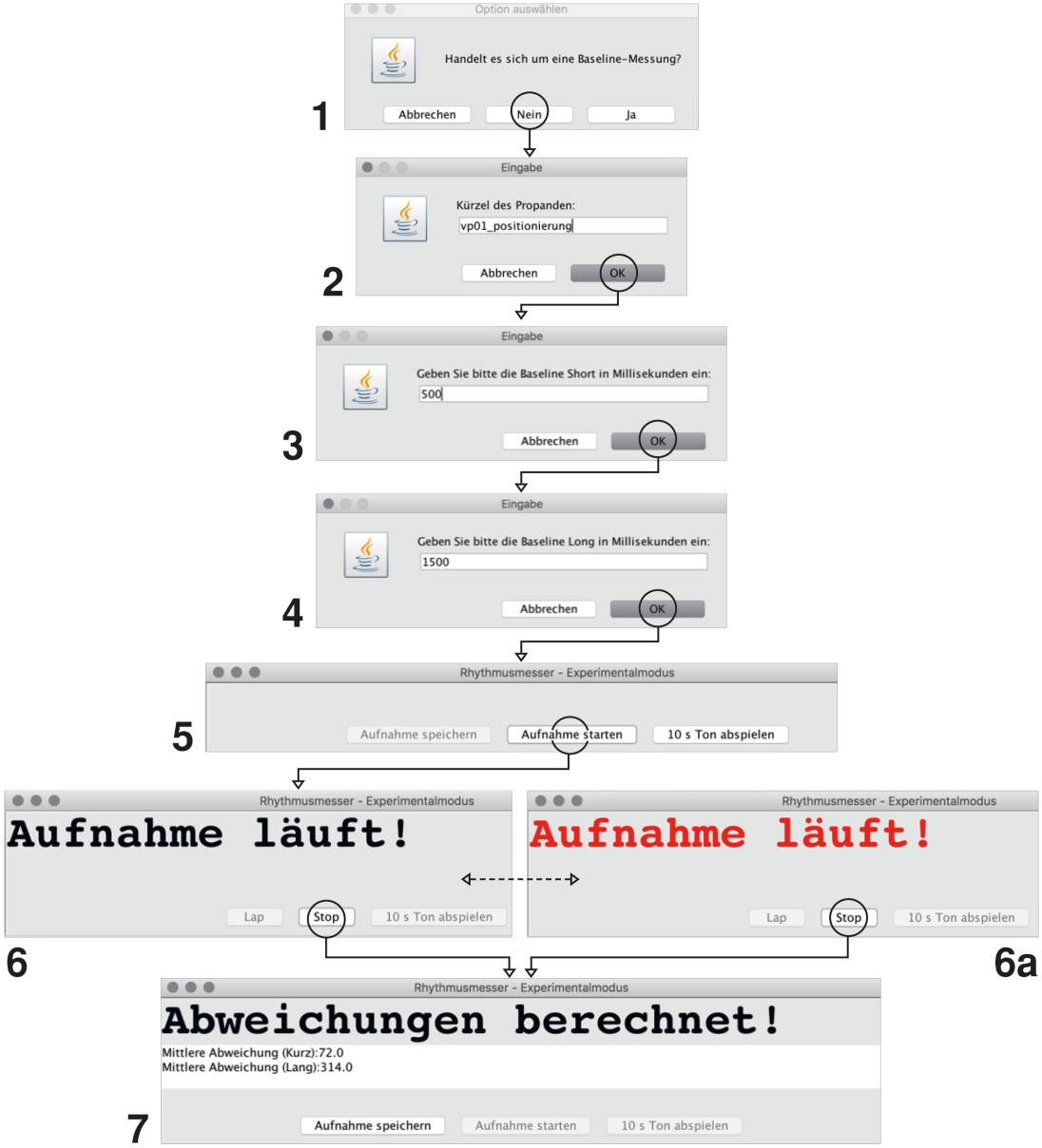


Abbildung 5.4. Flussdiagramm des Experimentalmodus der IntuiBeat-Software.

Nach der Spezifikation der erforderlichen Parameter für die Experimental-Messung wurden die Versuchspersonen zunächst vom Evaluator standardisiert mündlich instruiert, den gelernten Rhythmus parallel zur eigentlichen Systemnutzung zu klopfen. Der Evaluator wies sie im Rahmen dieser Instruktion speziell darauf hin, dass sie auf jeden Fall die Systemnutzung priorisieren sollen und ihre Performance bei der Systemnutzung wichtiger als die Rhythmusaufgabe sei. Auf diese Weise wurde vom Evaluator unter Berücksichtigung von Cain (2007) das Paradigma der Nebenaufgabe induziert (siehe Abschnitt 3.6). Das Hauptfenster des Experimental-Modus bot für diesen Zweck dem Evaluator die Möglichkeit, die

Rhythmusaufzeichnung zu starten, zu beenden, zu speichern oder daran zu erinnern (siehe Abbildung 5.4, Schritt 5).

Nachdem die Versuchspersonen dem Evaluator mündlich signalisiert hatten, dass seine standardisierte Instruktion verständlich war und sie nun mit der Systemnutzung beginnen möchten, spielte der Evaluator den Versuchspersonen noch einmal für 10 Sekunden den Rhythmus durch Klicken auf die Schaltfläche „10 s Ton abspielen“ als Audioaufzeichnung vor, um eine Inhibition des Rhythmus während der darauffolgenden Systemnutzung zu unterstützen (siehe Abbildung 5.4, Schritt 5). Daraufhin startete der Evaluator mithilfe der IntuiBeat-Software über die Schaltfläche „Aufnahme starten“ die Aufnahme. Nach Betätigung der Schaltfläche signalisierte die Software dem Evaluator die nun laufende Aufnahme mit den Worten „Aufnahme läuft“ (siehe Abbildung 5.4, Schritt 6). Wie bei der Baseline-Messung beschrieben, wurden bei der Experimental-Messung die gleichen Schwellenwerte wie bei der ursprünglichen Rhythmusmethode genutzt (siehe Abschnitt 4.1), um Eingaben dem kurzen und langen Rhythmusintervall zuordnen oder Eingaben verwerfen zu können (siehe Korbach et al., 2018; Park & Brünken, 2015). Der Evaluator wurde bei der Experimental-Messung ebenfalls von der Software über die Gültigkeit einer Rhythmuseinheit informiert, indem sich bei einer ungültigen Einheit die Worte „Aufnahme läuft“ rot einfärbten (siehe Abbildung 5.4, Schritt 6a). Die Versuchspersonen bekamen während der Experimental-Messung wie bei der Baseline-Messung nach jeder Eingabe von der Software akustisches Feedback.

Im Gegensatz zur Baseline-Messung stoppte die Rhythmusaufzeichnung jedoch bei der Experimental-Messung nicht automatisch, sondern musste vom Evaluator immer explizit mit einem Klick auf die Schaltfläche „Aufnahme stoppen“ beendet werden. Danach zeigte die Software dem Evaluator die durchschnittlichen Baseline-Abweichungen innerhalb der kurzen und langen Rhythmusintervalle an (siehe Abbildung super, Schritt 7). Abschließend speicherte der Evaluator die Rhythmusaufzeichnung durch Klicken auf die Schaltfläche „Aufnahme speichern“ (siehe Abbildung 5.4, Schritt 7). Im Anschluss wurden von der Software drei verschiedene Logs als CSV-Dateien generiert (d.h. Typ „[ID]_raw.csv“, Typ „[ID]_clean.csv“, Typ „[ID]_infos.csv“), die im selben Ordner auf dem Rechner des Evaluators abgelegt wurden (siehe Analyse-PC in Absatz „Apparatur“), in dem sich auch die JAR-Datei der Software befand. Die mithilfe des Experimental-Modus erzeugten Raw- und Clean-Dateien unterschieden sich von den korrespondierenden, mithilfe des Baseline-Modus erzeugten Dateien nur dadurch, dass hierin bei den jeweiligen Rhythmuseinheiten auch die Abweichung von der zuvor eingegebenen Baseline (d.h. Baseline des kurzen und Baseline des langen Rhythmusintervalls) vermerkt wurde. Eine Infos-Datei enthält den Mittelwert der Abweichungen von der kurzen Rhythmus-Baseline und den Mittelwert der Abweichungen von der langen Rhythmus-Baseline in Millisekunden. Die Software beendete sich nach der Speicherung der Aufnahme automatisch selbst. Es wurde sich dafür entschieden dem Evaluator auch bei der Experimental-Messung die Aufzeichnung explizit speichern zu lassen, da es ja auch hier zu Fehlern gekommen sein kann.

Summative Evaluation intuitiver Benutzung mit anderen Methoden

Als subjektive Methoden zur Erfassung des Gefühls von Flüssigkeit kamen die SEA-Skala (d.h. AV_SEA, siehe Eilers et al., 1986), der NASA-TLX in seiner ungerichteten

Raw-Version als NASA-RTLX (d.h. AV_NASA, siehe Byers, 1989) und der QUESI (d.h. AV_QUESI, siehe Naumann & Hurtienne, 2010) zum Einsatz. Bei der SEA-Skala, die nach jeder Aufgabe erhoben wurde, fungierte der gemittelte Testwert des Fragebogens aller Testaufgaben als Indikator (Wertebereich: 0 - 220). Beim NASA-RTLX, der nach jeder Aufgabe erhoben wurde, fungierte der gemittelte Gesamtwert (d.h. Mittelwert aller Skalen des Fragebogens) aller Testaufgaben als Indikator (Wertebereich: 0 - 100). Beim QUESI, der einmalig zur Beurteilung des Gesamtsystems erhoben wurde, fungierte der gemittelte Gesamtwert (d.h. Mittelwert über alle Skalen des Fragebogens) als Indikator (Wertebereich: 1 - 5). Alle subjektiven Methoden wurden in Papierform, und wie in Teilabschnitt 3.6.1 beschrieben, administriert.

Als objektive Methoden kam die Messung der Effektivität als Hauptaufgabenleistungsmaß zum Einsatz, wobei hier der Anteil korrekt abgeschlossener Aufgaben an den insgesamt durchgeführten Aufgaben (d.h. AV_Effektivität) als Indikator fungierte (siehe Teilabschnitt 3.6.2). Der Anteil korrekt abgeschlossener Aufgaben wurde durch den Evaluator bestimmt, indem dieser bei jeder Testaufgabe geprüft hat, ob die Aufgabenziele (z.B. „Wurde der Stuhl wirklich um 180 Grad gedreht?“) von den Versuchspersonen erreicht wurden oder nicht. Um entscheiden zu können, ob die Aufgabe effektiv bearbeitet wurde, mussten der Evaluator für jede Testaufgabe und bearbeitete Software a priori festlegen, welche möglichen Lösungswege für diese Testaufgabe zielführend sind. Im Anschluss erfolgte durch den Evaluator eine Umkodierung der künstlichen dichotomen Variable (d.h. „0“ wenn nicht gelöst; „100“ wenn gelöst), um trotz des kategorialen Datenniveaus später auf metrischen Datenniveaus rechnen zu können. Ein solches Vorgehen stellt einen üblichen statistischen Trick bei künstlichen dichotomen Variablen dar, welchen ein feiner skaliertes Merkmal zugrunde liegt (Bortz & Schuster, 2011; MacCallum, Zhang, Preacher, & Rucker, 2002).

Zusätzlich kam noch die CHAI-Methode von Reinhardt et al. (2018) als objektives konvergentes Quasi-Außenkriterium zum Einsatz (d.h. AV_CHAI), welche als objektiver Benchmark der summativen Evaluation in Abschnitt 3.4 identifiziert werden konnte. Als Indikator fungierten hierbei der Anteil aller intuitiven Mausklicks während der gesamten Systemnutzung an der Gesamtzahl der getätigten Mausklicks. Es wurden Mausklicks zur Identifikation von Interaktionen gewählt, da alle Testaufgaben des ersten und auch der übrigen Experimente so gestaltet waren, dass sie sich ausschließlich mit der Maus haben lösen lassen. Dementsprechend stellte jeder Mausklick (d.h. Klick auf die linke oder rechte Maustaste) eine Interaktion mit dem System dar. Um den Anteil aller intuitiven Mausklicks bei der Systemnutzung bestimmen zu können, beurteilte der Evaluator die Systemnutzung der Versuchspersonen retrospektiv mithilfe des in Teilabschnitt 3.6.2 vorgestellten CHAI-Beurteilungsschemas anhand der Videoaufzeichnungen, in denen die Versuchspersonen die beiden CUIs zur Lösung der Aufgaben benutzten und bei dieser Nutzung zusätzlich den für IntuiBeat-S benötigten Rhythmus klopfen mussten.

Wie bei der ursprünglichen Rhythmusmethode handelt es sich auch bei IntuiBeat-S um eine auf Inhibitionsprozessen basierende Zweitaufgabe, welche für die Erfassung von mentaler Beanspruchung im Rahmen eines Zweitaufgabenparadigmas die nötige Interferenz bei minimaler Intrusion erzeugen kann (siehe Abschnitt 4.1). Dementsprechend wurde sich beim beschriebenen Experiment für die Durchführung der CHAI-Methode auf Basis der Videoaufzeichnungen mit paralleler Rhythmusausführung entschieden, da damit

theoretisch keine große Intrusion zu befürchten war. Mithilfe des freien, quelloffenen Mediaplayers VLC von VideoLAN (2017) wurde vom Verfasser dieser Dissertation jeder in den Videoaufzeichnungen enthaltene Mausklick, der zur Interaktion mit dem System von den Versuchspersonen getätigt wurde, auf Basis des CHAI-Beurteilungsschemas bewertet. Neben dem eigentlichen CHAI-Beurteilungsschema veröffentlichten Reinhardt et al. (2018) auch einen Entscheidungsbaum für die Anwendung, der in Abbildung 5.5 dargestellt ist und für die Bestimmung des Anteils aller intuitiven Mausklicks im Rahmen des ersten Experiments und der beiden Folgeexperimente Anwendung fand.

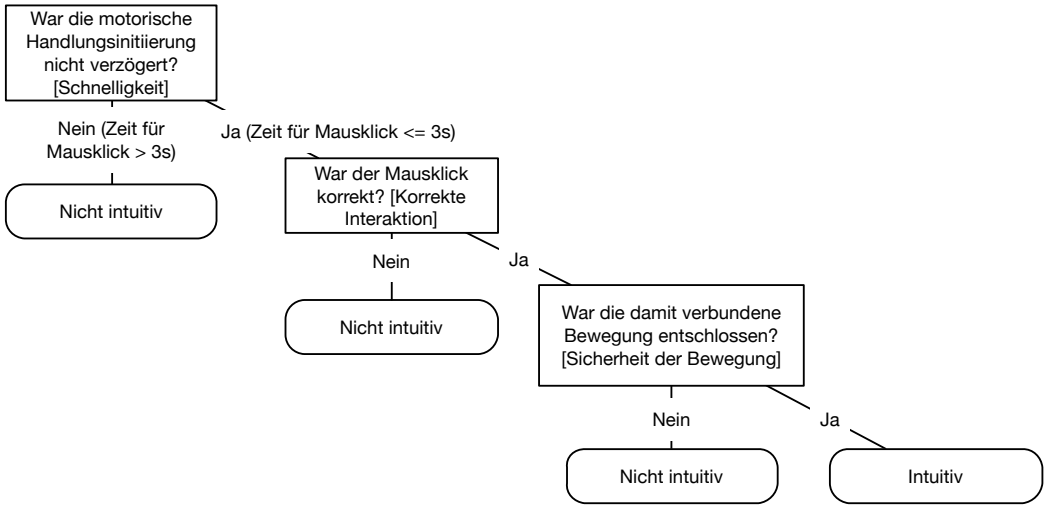


Abbildung 5.5. Entscheidungsbaum, um mithilfe der CHAI-Methode zwischen *intuitiven* und *nicht intuitiven* Interaktionen differenzieren zu können (Reinhardt, Kuge, & Hurtienne, 2018). Wurde eines der Kriterien verletzt, wurde der Klick als nicht intuitiv aufgefasst und mit der Beurteilung des nächsten Klicks fortgefahren.

Da eine intuitive Benutzung bei vielen Interaktionen bereits durch eine mangelnde Effektivität und eine langsame Handlungsinitiierung (d.h. geringe zeitliche Effizienz bei der Informationsverarbeitung) ausgeschlossen werden kann, stellen diese beiden Kriterien die ersten beiden Entscheidungen des Entscheidungsbaums dar. Die Sicherheit der Bewegung während der Interaktion mit dem System wurde dementsprechend, wie von Reinhardt et al. (2018) vorgeschlagen, immer als letztes Kriterium vom Evaluator beurteilt. Um mithilfe des Entscheidungsbaums die Videoaufzeichnungen der Testaufgaben kodieren zu können, analysierte der Evaluator unter Berücksichtigung der Empfehlungen von Reinhardt et al. (2018) jede Videoaufzeichnung sukzessive anhand von drei Sekunden langen Zeitfenstern in Abhängigkeit des letzten erfolgten Mausklicks. Zeigte sich innerhalb von drei Sekunden keine motorische Handlungsinitiierung (d.h. Klick mit der Maus) seitens des Versuchsteilnehmers wurde das Kriterium Schnelligkeit von diesem nicht erfüllt und die Interaktion wurde dementsprechend vom Evaluator als nicht intuitiv gewertet (siehe Abbildung 5.5, Schritt 1). Es wird an dieser Stelle bewusst von motorischer Handlungsinitiierung gesprochen, da der Nutzer ja noch mit der eigentlichen Aufgabe (z.B. sich den Inhalt einer Mail überlegen, die er gerne mit dem System verfassen möchte) beschäftigt sein könnte und man ja nicht diese vom System unabhängige Vorarbeit, sondern die Intuitivität der Interaktion mit dem System bewerten möchte.

Fand ein Mausklick innerhalb dieses Zeitfensters statt und erfolgte dadurch die motorische Handlungsinitiierung, überprüfte der Evaluator als nächstes, ob dieser Mausklick korrekt war oder nicht (siehe Abbildung 5.5, Schritt 2). Mausklicks werden laut Reinhardt et al. (2018) dann als korrekt gewertet, wenn der Mausklick den Versuchsteilnehmer näher an sein zuvor gesetztes Handlungsziel (z.B. Position einer Lampe im Raum verändern) bringen konnte, ohne dass der Versuchsteilnehmer dabei explizit Hilfe vom Evaluator oder vom System angefordert hat. Um entscheiden zu können, ob der Mausklick wirklich für die Interaktion zielführend war, musste der Evaluator vor der Analyse der Mausklicks mit der CHAI-Methode a priori für jede Testaufgabe festlegen, welche möglichen Lösungswege für diese Testaufgabe zielführend sind. Falls der Mausklick unter Berücksichtigung dieser Lösungswege als effektiv eingestuft werden konnte, fuhr der Evaluator mit dem nächsten Schritt im Entscheidungsbaum fort. Konnte der Mausklick hingegen vom Evaluator nicht als effektiv eingestuft werden, wurde dieser auch nicht als intuitiv gewertet (siehe Reinhardt et al., 2018).

Konnte ein Mausklick vom Evaluator schließlich als schnell und effektiv eingestuft werden, konnte von ihm auch das letzte Kriterium der Sicherheit der Bewegung beurteilt werden (siehe Abbildung 5.5, Schritt 3). Wurde die Mausbewegung bis zum Mausklick vom Evaluator als unsicher und indirekt (d.h. zögerliche, unterbrechende, kreisende Mausbewegungen) eingestuft, wurde der Mausklick vom Evaluator als nicht intuitiv gewertet, da von der Interaktion somit nicht alle Anforderungen an eine intuitive Interaktion erfüllt wurden. Führte die Versuchsperson hingegen ihren Mauscursor ohne größere Abweichungen direkt zum Ziel, wurde das Kriterium der Sicherheit der Bewegung vom Evaluator als gegeben eingestuft und auch der Mausklick entsprechend als intuitiv angesehen (Reinhardt et al., 2018). Nachdem ein Mausklick mithilfe des beschriebenen Entscheidungsbaums (siehe Abbildung 5.5) als intuitiv oder nicht intuitiv klassifiziert werden konnte, hielt der Evaluator dieses Gesamtergebnis in einer CSV-Datei fest.

Das oben geschilderte Vorgehen wurde dann für jedes Zeitfenster von drei Sekunden der analysierten Testaufgaben wiederholt. Nachdem auf diese Weise alle Videoaufzeichnungen der Testaufgaben mithilfe der CHAI-Methode evaluiert wurden, konnte der Anteil an Intuitiv-Klicks im Verhältnis zu allen getätigten Klicks auf Aufgaben-Ebene berechnet werden. Durch Mittelwertbildung dieser Werte konnte der Anteil aller intuitiven Mausklicks während der gesamten Systemnutzung an der Gesamtzahl der getätigten Mausklicks für jeden Versuchsteilnehmer ermittelt werden. Aufgrund der Tatsache, dass die Meta-Evaluation der CHAI-Methode ein κ von .61 aufwies, was laut Field (2017) und Landis und Koch (1977) einer substantiellen Übereinstimmung entspricht, wurde im Rahmen des ersten Experiments und der Folgeexperimente die CHAI-Methode lediglich alleine von einem Evaluator durchgeführt. Bei diesem handelte es sich um den Verfasser dieser Dissertation und auch um einen Urheber der CHAI-Methode (siehe Reinhardt et al., 2018), der deswegen auch die größte Vorerfahrung mit der Anwendung der Methode aufweisen konnte. Auf eine erneute Berechnung der Beurteilerübereinstimmung wurde dementsprechend an dieser Stelle verzichtet, da nicht die Meta-Evaluation der CHAI-Methode der Gegenstand dieser Arbeit war.

Summative Evaluation von physischer und zeitlicher Effizienz

Neben den vorgestellten konvergenten Evaluationsmethoden, die zur Überprüfung der konvergenten Validität von IntuiBeat-S genutzt wurden, kamen im Rahmen des beschriebenen Experiments noch zwei weitere objektive Methoden zum Einsatz. Diese erfasst die physische Effizienz der Systemnutzung (d.h. AV_PhysischeEffizienz) und die zeitliche Effizienz der motorischen Handlungsdurchführung (d.h. AV_ZeitlicheEffizienz). Diese beiden Methoden fungierten damit als Quasi-Außenkriterien für die Überprüfung der divergenten Validität (Hypothesen H3.A und H3.B). Als Indikator der physischen Effizienz wurde die Anzahl der benötigten Klicks bei der Systeminteraktion verwendet, so wie sie aufgrund der CHAI-Methode vorlagen. Die Zeit der motorischen Systemnutzung in Sekunden (d.h. die zur Bearbeitung aller Testaufgaben aufsummierte motorische Gesamthandlungszeit, in der der Nutzer die Maus bewegte und sie nicht stillstand) fungierte als Indikator für die zeitliche Effizienz der motorischen Handlungsausführung und wurde manuell anhand der, für die CHAI-Methode genutzten, Videoaufzeichnungen bestimmt (siehe Teilabschnitt 2.1.1).

Im Rahmen des beschriebenen Experiments wurden neben den vorgestellten Evaluationsmethoden noch demographische Daten und verschiedene Kontrollvariablen mithilfe der Umfragesoftware LimeSurvey (LimeSurvey Development Team, 2015) erhoben.

Demographische Variablen

Als soziodemografische Daten wurden von den Versuchspersonen Alter, Geschlecht und Studiengang erhoben.

Vorerfahrung bei der Nutzung von CUIs

Die *Vorerfahrung bei der Nutzung von CUIs* wurde als Kontrollvariable (d.h. KV_Vorerfahrung) mit einem für 3D-CUIs konstruierten TFQ erhoben. Als Items für diesen Fragebogen wurden die CUIs „Fusion 360“, „AutoCAD“, „SketchUp“, „SolidWorks“ und damit Beispiele für bekannte CUIs verwendet. Darüber hinaus wurde „Andere CAD-Software?“ als Wildcard genutzt, um die Eingabe eines beliebigen bekannten CUI zu gestatten. Es wurde sich für die Verwendung des Begriffs „CAD“ anstelle von „CUI“ entschieden, da dieser Begriff in der genutzten Stichprobe wahrscheinlich geläufiger ist und damit keiner Einführung bedürfte. Versuchspersonen beurteilten die Items dann lediglich bezüglich deren Nutzungshäufigkeit (Häufigkeitsdimension: „Wie häufig verwenden Sie die folgenden CAD-Anwendungen?“; Wertebereich: 0 - 6). Es wurde für die Operationalisierung der Vorerfahrung die Häufigkeitsdimension des TFQ verwendet, obwohl dessen wissenschaftliche Güte nicht über die Inhaltsvalidität hinaus nachgewiesen werden konnte (siehe Abschnitt 3.4), da als Alternative nur die Verwendung einer eigenen Skala in Frage gekommen wäre. Da kein einheitlicher TFQ existiert und dieser jeweils für die getesteten Softwareanwendungen erstellt werden musste, ist der in diesem Experiment genutzte TFQ vollständig in Anhang B.1.2 dieser Arbeit zu finden.

Im Rahmen des ersten Experiments hat man sich gegen eine Beurteilung bezüglich des genutzten Funktionsumfangs und damit nur für die Verwendung einer Beurteilungsdimension entschieden, da bei der verwendeten Stichprobe (d.h. Studierende der Medienkommunikation und der Mensch-Computer-Systeme) davon auszugehen war, dass aufgrund der geringen Relevanz von CUIs im Studium wenig Varianz bereits auf der Häufigkeitsdimension zu erwarten ist, und demzufolge bezüglich des Funktionsumfangs noch weniger. Der Mittelwert der Häufigkeitsdimension fungierte dementsprechend als Operationalisierung der *Vorerfahrung bei der Nutzung von CUIs* im Rahmen des ersten Experiments. Es wurde sich gegen die von Blackler (2006) und in Abschnitt 3.4 erläuterte Summenberechnung für eine Mittelwertbildung entschieden, da die Operationalisierung der Vorerfahrung als Mittelwert logischer erscheint, den sonst von der Länge abhängigen TFQ-Wert normiert und Vergleiche zwischen den Experimenten zulässt.

Rhythmische Wahrnehmung

Da es trotz individuell erhobener Baselines sein kann, dass die musikalischen Fähigkeiten der Versuchsteilnehmer, insbesondere deren Rhythmusgefühl, die wissenschaftliche Güte von IntuiBeat-S beeinflussen können, wurden diese Fähigkeiten in Form einer weiteren Kontrollvariable (d.h. KV_Rhythmik) berücksichtigt. Für die Erfassung von musikalischen Fähigkeiten wird oftmals der Grad der musikalischen Ausbildung herangezogen, was aber dazu führt, dass sich bei Personen ohne musikalische Ausbildung musikalische Fähigkeiten nicht erfassen lassen (Law & Zentner, 2012). Im Gegensatz dazu bietet der PROMS (d.h. Profile of Music Perception Skills) als Methode die Möglichkeit die musikalische Kompetenz der Versuchspersonen objektiv in verschiedenen Bereichen (d.h. Melodie, Rhythmus, Rhythmus-zu-Melodie, Stimmung, Akzent, Instrumente, Tempo, Tonhöhe, Lautstärke) und ohne die Anforderung an eine musikalische Ausbildung zu erfassen (Law & Zentner, 2012; Zentner & Strauss, 2017). Sie eignet sich daher zur Einschätzung der musikalischen Wahrnehmung der Probanden im Rahmen der hier beschriebenen Studie. Laut Kunert, Willems und Hagoort (2016) steht die musikalische Wahrnehmung in direkten Zusammenhang mit der tatsächlichen musikalischen Performance und gibt demzufolge Aufschluss darüber wie die unterschiedlichen Wahrnehmungsbereiche zur musikalischen Kompetenz von Personen beitragen. Die musikalische Wahrnehmung der Versuchspersonen wird hierzu mithilfe des PROMS bezüglich jedes musikalischen Wahrnehmungsbereichs anhand einer Reihe von Probe- und Experimentaldurchläufen (d.h. Trials) bewertet.

Jeder Durchlauf läuft dabei gleich ab und enthält drei Audiobeispiele, anhand derer sich ein bestimmter Wahrnehmungsbereich von den Versuchspersonen beurteilen lässt (siehe Abbildung 5.6). Den Versuchspersonen werden bei diesem Verfahren zunächst zwei identische Audiobeispiele nacheinander und nach einer kurzen Pause ein drittes Audiobeispiel vorgespielt. Die Versuchspersonen müssen dann einschätzen, ob sich die Vergleichs-Audioaufzeichnung hinsichtlich des Wahrnehmungsbereichs (z.B. Melodie) von den zuvor dargebotenen Audiobeispielen unterscheidet. Für diese Einschätzung stehen den Versuchspersonen fünf Antwortmöglichkeiten zur Auswahl (d.h. „definitiv gleich“, „wahrscheinlich gleich“, „wahrscheinlich verschieden“, „definitiv verschieden“, „Ich weiß es nicht“). Zur Berechnung eines Scores für einen Wahrnehmungsbereich werden alle korrekten „definitiven“ Antworten mit einem Punkt und alle korrekten „wahrscheinlichen“ Antworten innerhalb

der Experimentaldurchläufe mit einem halben Punkt bewertet. Inkorrekte Antworten (d.h. egal, ob „definitiv“ oder „wahrscheinlich“) oder Enthaltungen (d.h. „Ich weiß es nicht“) innerhalb der Experimentaldurchläufe werden mit keinem Punkt bewertet. Der Score des jeweiligen Wahrnehmungsbereichs ergibt sich dann durch Aufsummieren der Ergebnisse der dazugehörigen Experimentaldurchläufe. Der Gesamtscore zur Erfassung der musikalischen Wahrnehmung wird dann durch Berechnung des Mittelwerts (auch Summenbildung möglich, wobei im Rahmen des Experiments zur Beibehaltung der Konsistenz gegenüber anderen Maßen gemittelt wurde) der einzelnen Scores der Wahrnehmungsbereiche gebildet. Handelt es sich bei dem Durchlauf nicht um einen Experimentaldurchlauf, sondern um einen Probedurchlauf, gibt der PROMS den Versuchspersonen Feedback über ihre Einschätzung und teilt ihnen bei einer inkorrekten Antwort auch die korrekte Einschätzung mit. Das Vorgehen und die Auswertung wurde aus der Veröffentlichung von Law und Zentner (2012) entnommen und für weitere Informationen soll an dieser Stelle auch auf entsprechende Arbeit verwiesen werden.

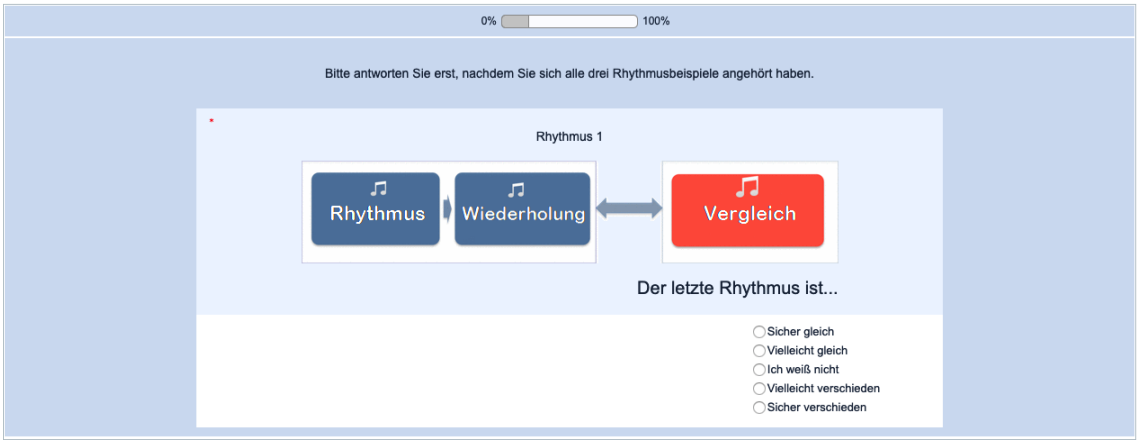


Abbildung 5.6. Experimentaldurchlauf zur Beurteilung des Wahrnehmungsbereichs Rhythmus innerhalb des PROMS (siehe Law & Zentner, 2012).

Der PROMS existiert, neben einer Langfassung (Law & Zentner, 2012), auch in einer Kurzversion, die sich von der Langversion dadurch unterscheidet, dass sie weniger Durchläufe (d.h. die Kurzversion benötigt pro Wahrnehmungsbereich um die 10 Durchläufe, was der Hälfte der Langversion entspricht) nutzt, was aber nur marginal auf Kosten der wissenschaftlichen Güte und den damit verbundenen psychometrischen Gütekriterien geht (Zentner & Strauss, 2017). Die wissenschaftliche Güte des PROMS (z.B. Kunert et al., 2016; Law & Zentner, 2012) und des PROMS-S (z.B. Zentner & Strauss, 2017) wurde schon anhand mehrerer Studien nachgewiesen. Da auch diese Kurzversion PROMS-S mit 30 Minuten immer noch eine verhältnismäßig lange Administrationszeit hat und bei der Anwendung des bei IntuiBeat-S verwendeten Rhythmus überwiegend musikalische Fähigkeiten benötigt werden, die mit dem allgemeinen Rhythmusgefühl der Versuchsteilnehmer zu tun haben, wurde lediglich ein modularer PROMS-S im Rahmen des Experiments zur Erfassung der rhythmischen Wahrnehmung genutzt. Dieser fragte lediglich die damit zusammenhängenden Wahrnehmungsbereiche (d.h. Rhythmus: acht Durchläufe, Rhythmus-zu-Melodie: acht Durchläufe, Akzent: zehn Durchläufe) der Versuchspersonen ab (siehe Zentner & Strauss, 2017). Der gemittelte Gesamtscore über diese drei Wahrnehmungsbereiche (Wertebereich: 0 - 8.67) kam im Rahmen des Experiments zur Erfassung der

rhythmischen Wahrnehmung der Versuchspersonen als Operationalisierung zum Einsatz. Die wissenschaftlichen Güte eines solchen modularen PROMS-S konnte ebenfalls bereits nachgewiesen werden (Zentner & Strauss, 2017).

Apparatur

Die Versuchspersonen bearbeiteten die Testaufgaben mit den CUIs auf einem MSI GT62VR Gaming Notebook (Windows 10 Home, 2.8 GHz Quad-Core Intel i7, 16 GB Arbeitsspeicher, 8 GB NVIDIA GeForce GTX 1070 Grafikkarte), der somit als Versuchspersonen-PC diente (siehe Abbildung 5.7). Es wurde ein höhenverstellbarer Drehstuhl als Sitzgelegenheit für die Versuchsperson verwendet, um dieser die Rhythmuseingabe über das USB-Fußpedal so ergonomisch wie möglich zu erlauben. Als Eingabegeräte fungierten dabei nicht die eingebaute Tastatur und das Trackpad des Notebooks, sondern eine externe kabelgebundene Maus M-UAV-DEL8 von Dell und eine externe kabelgebundene Tastatur K200 von Logitech. Um die rhythmische Wahrnehmung der Versuchspersonen mithilfe des PROMS testen zu können, wurde noch ein Studiokopfhörer HD-668 B von Superlux an den Versuchspersonen-PC angeschlossen. Als Bildschirm für den Versuchspersonen-PC kam ein dedizierter 24 Zoll LCD-Bildschirm FLATRON E2411 von LG (Auflösung 1920 x 1080 Pixel) zum Einsatz. Da bei CUIs viele Elemente auf dem Bildschirm angezeigt werden müssen und deswegen in der Praxis entsprechend große Monitore zum Einsatz kommen, wurde sich im Rahmen des beschriebenen ersten Experiments entschieden, dass die Meta-Evaluation von IntuiBeat-S in einem vergleichbaren Setting erfolgen muss. Der Abstand zwischen den Versuchspersonen und dem Bildschirm betrug hier circa 70 cm. Der eingebaute 15 Zoll Bildschirm des Gaming Notebooks wurde während der Erhebung vollständig abgedunkelt, um Ablenkungen zu vermeiden.

Die Versuchspersonen nutzten auf diesem Notebook neben den CUIs auch die Umfragesoftware LimeSurvey von LimeSurvey Development Team (2015), durch die von ihnen die Kontrollvariablen und demographischen Daten der Studie erfragt wurden. Als Browser kam hier Google Chrome (Version 62.0.320) zum Einsatz. Die als Faktorstufen der unabhängigen Variable fungierenden CUIs wurden vom Evaluator im Vorfeld des Experiments auf dem Versuchspersonen-PC installiert. Während der Bearbeitung der Testaufgaben mit den beiden CUIs wurde mit der Usability-Testing Software *Morae* von TechSmith (2004) der Bildschirm der Versuchspersonen für die Durchführung der Meta-Evaluation aufgezeichnet. Morae ist eine Software-Suite, die eigentlich drei verschiedene Softwarelösungen enthält: Recorder, Observer und Manager.

Der Recorder muss für eine Aufzeichnung des Bildschirms auf dem Versuchspersonen-PC installiert sein. Neben dem Bildschirm lassen sich damit auch Audio, Video über die Webcam des MSI-Notebooks und die Eingaben der Versuchsperson (z.B. Tastendrücken, Mausbewegungen) aufzeichnen. Im Rahmen des berichteten Experiments wurden damit nur der Bildschirm und Mausbewegungen der Versuchsperson aufgezeichnet, da nur diese für die summative Meta-Evaluation von IntuiBeat-S relevant waren. Mithilfe des Observers konnte der Evaluator, die vom Recorder aufgezeichneten Parameter in Echtzeit an einem anderen Rechner (d.h. Evaluatoren-PC, siehe Abbildung 5.7) verfolgen und die Aufzeichnung von diesem Rechner aus steuern, nachdem er Recorder und Observer durch Eingabe einer IP-Adresse auf Observer-Seite verbunden hatte (z.B. Starten oder Stoppen

der Bildschirmaufzeichnung). Die Software gestattet es zudem, an interessanten Stellen (z.B. auftretendes Nutzungsproblem) innerhalb der Aufzeichnung Marker zu setzen, was vom Evaluator für die Erhebung von qualitativen Daten im Rahmen des Experiments zwar genutzt wurde, jedoch für die summative Meta-Evaluation von IntuiBeat-S irrelevant ist. Die Observer-Software wurde auf dem Evaluatoren-PC installiert, der sich direkt hinter dem Arbeitsplatz der Versuchsperson bzw. des Versuchspersonen-PCs befand und vom Evaluator während der Nutzertests genutzt wurde (siehe Abbildung 5.7). Bei dem Rechner handelt es sich um einen Fujitsu Esprimo P710 (Windows 10 Home, 3.4 GHz Dual-Core Intel i3, 4 GB Arbeitsspeicher, Intel HD Graphics 2500 Grafikkarte), welcher mithilfe einer externen kabelgebundenen Maus M-U0026 von Fujitsu und einer Tastatur K200 von Logitech vom Evaluator gesteuert wurde. Als Bildschirm kam ebenfalls ein 24 Zoll LCD-Bildschirm FLATRON E2411 von LG (Auflösung 1920 x 1080 Pixel) zum Einsatz. Der Abstand zwischen dem Evaluator und dem Bildschirm betrug wie bei den Versuchspersonen ebenfalls circa 70 cm.

Abbildung 5.7. Apparatur des ersten Experiments, bestehend aus Versuchspersonen-PC (hinten), Analyse-PC (vorne links) und Evaluatoren-PC (vorne rechts).

Um die vom Recorder aufgezeichneten Dateien öffnen, diese bearbeiten und für die Datenanalyse vorbereiten zu können, wird als letztes Tool des Softwarepakets Morae noch der Manager benötigt. Da es das Lizenzmodell von TechSmith, dem Herstellers von Morae, zwar gestattet beliebig viele Recorder und Observer zu installieren, aber man pro Lizenz nur einen Manager auf einem beliebigen System installieren darf, wurde als Analyse-PC (siehe Abbildung 5.7) ein dediziertes Apple Macbook Air (Windows 7 Home, 2.3 GHz Dual-Core Intel i7, 8GB Arbeitsspeicher, 1536 MB Intel HD Graphics 6000 Grafikkarte, Auflösung 1280 x 800 Pixel) gewählt, um dieses auch für andere Studien verfügbar zu haben. Aufgrund der Tatsache, dass alle Komponenten von Morae nur mit Windows funktionieren, wurde auf dem Macbook Windows 7 mit Apple Boot Camp installiert. Neben der Manager-Software wurde auf diesem Notebook auch, die für die Anwendung der getesteten objektiven konvergenten Evaluationsmethoden benötigten, Softwarepakete VLC Mediaplayer von VideoLAN (2017) (CHAI-Methode) und die IntuiBeat-Software (IntuiBeat-S) installiert.

Es wurde für die Anwendung der Evaluationsmethoden ein von der eigentlichen Aufzeichnung unabhängiger Rechner auch deshalb gewählt, weil für IntuiBeat-S der Rhythmus parallel zur Systemnutzung aufgezeichnet werden musste und man keine hohen Verzögerungszeiten der Hardware riskieren wollte, wenn die Videoaufzeichnung und die Rhythmusaufzeichnung auf demselben System mithilfe der IntuiBeat-Software stattfinden. Für die Rhythmusaufzeichnung wurde an den Analyse-PC das USB-Fußpedal von Sodial per USB 2.0 angeschlossen und unter dem Tisch positioniert, auf dem sich der Versuchspersonen-PC befand. Da die IntuiBeat-Software in Java geschrieben wurde, wurde für die Ausführung ein Java Runtime Environment benötigt, das zu diesem Zweck in Version 1.8.0_111 auf dem Analyse-PC installiert wurde. Die Aufzeichnungsdateien wurden zwischen den Systemen (d.h. Versuchspersonen-PC, Evaluatoren-PC und Analyse-PC) über ein gemeinsames Netzlaufwerk ausgetauscht.

5.1.4.4 Versuchsdurchführung

Die Rekrutierung der Versuchspersonen erfolgte über das Probanden-System des Instituts für Mensch-Computer-Medien der Universität Würzburg. Das Experiment wurde darin als „Nutzertest verschiedener Software-Anwendungen“ beworben (siehe Anhang A.1). In der Beschreibung wurde potentiellen Versuchsteilnehmern mitgeteilt, dass das Ziel des Experiments darin besteht, Probleme und positive Aspekte verschiedener Software-Anwendungen ausfindig zu machen. Alle angebotenen Termine waren Einzeltermine, da immer nur eine Person gleichzeitig getestet wurde.

Das Experiment wurde in einem kleinen, reizabgeschirmten, stimulusarmen Labor des Lehrstuhls für Psychologische Ergonomie der Universität Würzburg durchgeführt. Der Geräuschpegel von außen war über den gesamten Erhebungszeitraum konstant niedrig. Bevor die Versuchspersonen das Testlabor betreten durften, wurden sie vom Evaluator mündlich aufgefordert, ihre Mobiltelefone in den Flugmodus zu schalten, um eine Ablenkung durch diese während der Untersuchung zu vermeiden. Die eigentliche Datenerhebung der Studie unterteilte sich in drei Phasen (d.h. Vorphase, Hauptphase, Abschlussphase), die insgesamt circa 60 Minuten in Anspruch nahmen.

Vorphase (20 Minuten)

Die Vorphase des Experiments begann mit der standardisierten mündlichen Begrüßung der Versuchsteilnehmer durch den Evaluator. Hierbei informierte der Evaluator zunächst standardisiert mündlich über den Ablauf der Studie, die freiwillige Teilnahme und die vertrauliche Behandlung der erhobenen Daten. Danach ließ der Evaluator die Versuchspersonen eine schriftliche Einverständniserklärung (siehe Anhang A.4) unterzeichnen und öffnete daraufhin die Umfragesoftware *LimeSurvey*. Versuchspersonen füllten im Anschluss den demographischen Fragebogen und den TFQ aus. Danach wurden sie vom Evaluator standardisiert mündlich instruiert, bei der Bearbeitung des nun folgenden Tests ihrer rhythmischen Wahrnehmung über den PROMS den Bildschirminstruktionen in LimeSurvey zu folgen und dazu die Kopfhörer aufzusetzen (siehe Teilabschnitt 5.1.4.3). Der Evaluator wies hierbei die Versuchspersonen außerdem darauf hin, dass sie bei etwaigen Unstimmigkeiten Fragen an den Evaluator richten können. Die Bearbeitung des PROMS sollte aber

trotzdem stillschweigend erfolgen. Der PROMS begann mit einem Kalibrierungstest der Soundwiedergabe, der bei den Versuchspersonen abfragte, inwiefern die Audiovideogabe des Systems problemlos verständlich ist. Anschließend bearbeiteten die Versuchspersonen den modularen PROMS (d.h. Wahrnehmungsbereiche: Rhythmus, Rhythmus-zu-Melodie und Akzent) standardmäßig, so wie dies ausführlich von Zentner und Strauss (2017) beschrieben wurde und für weitere Informationen deswegen an dieser Stelle darauf verwiesen wird.

Abschließend wurde den Versuchspersonen ihr Gesamtscore inklusive einer allgemeinen Interpretation darüber, was höhere und niedrigere Testergebnisse in den Wahrnehmungsbereichen für ihre rhythmische Wahrnehmung bedeuten, über LimeSurvey präsentiert. Nachdem sich die Versuchspersonen ihr Testergebnis kurz durchgelesen hatten, erfolgte im Anschluss eine standardisierte mündliche Instruktion durch den Evaluator für die folgende Hauptphase der Untersuchung.

Hauptphase (35 Minuten)

Im Rahmen der standardisierten mündlichen Instruktion wurde den Versuchspersonen zunächst der genaue Ablauf des eigentlichen Nutzertests mitgeteilt und ihnen dabei erläutert, dass das Ziel eines solchen Tests darin bestehe, sowohl Probleme des Systems zu identifizieren, als auch das Ausmaß intuitiver Benutzung der CUIs zu bewerten. An dieser Stelle teilte der Evaluator den Versuchspersonen ebenfalls mit, dass die intuitive Benutzung neben anderen Methoden (z.B. Fragebögen) auch mithilfe einer Rhythmuszweitaufgabe im Rahmen der Hauptphase gemessen werden soll. Der Evaluator erklärte den Versuchspersonen daraufhin, dass Rhythmusabweichungen dafür nur genutzt werden könnten, wenn diese in Abhängigkeit einer eigenen Rhythmus-Baseline gemessen würden, da sie nur so unabhängig von individuellen Unterschieden im Rhythmusgefühl und Musikalität sein könnten. Der Evaluator instruierte im Anschluss die Versuchsteilnehmer standardisiert mündlich, dass im nächsten Schritt die Baseline für die eigentliche Systemnutzung erhoben wird. Nachdem der Evaluator die Position des USB-Fußpedals so adjustiert hatte, dass die Versuchspersonen angenehm einen Rhythmus klopfen konnten, verkabelte er das USB-Fußpedal mit dem Analyse-PC und startete daraufhin die IntuiBeat-Software. Das Fußpedal war nur solange in den Analyse-PC einsteckt wie es für die eigentliche Rhythmusaufzeichnung (d.h. nicht für Fragebögen, Instruktion, Vorbereitung der Software, Interview oder Ähnliches) erforderlich war. Damit sollten etwaige Fehleingaben der Versuchspersonen (z.B. Versuchsperson drückt das Fußpedal während sich der Evaluator im Eingabefeld für das Versuchspersonenkürzel befindet, was eine ungewollte „0“ ins Eingabefeld einfügt) vermieden werden. Im Anschluss führte der Evaluator die Baseline-Messung nach der Vorgehensweise durch, wie sie bei der Beschreibung von IntuiBeat-S im vorigen Teilabschnitt 5.1.4.3 in Form einer generellen Methodenbeschreibung in Absatz „Funktionsweise des Baseline-Modus der IntuiBeat-Software und Beschreibung des Programmablaufs“ dargestellt wurde.

Im Anschluss an die Baseline-Messung stellte der Evaluator den Versuchsteilnehmern mündlich ein kurzes, fiktives Raumplanungsszenario vor, welches als Rahmenhandlung für die mit den CUIs zu erledigten Testaufgaben dienen sollte. Daraufhin händigte der Evaluator den Versuchspersonen eine Beschreibung des Szenarios und randomisiert die

erste Testaufgabe in Papierform aus. Auf diese Weise wurden die Versuchspersonen auch den Versuchsbedingungen *weniger intuitiv benutzbare Software* oder *stärker intuitive benutzbare Software* zugewiesen und diese Zuweisung bezüglich der unterschiedlich intuitiv benutzbaren Software (d.h. unabhängige Variable) über alle Versuchspersonen hinweg ausbalanciert. Der Evaluator öffnete anschließend die entsprechende Software (d.h. weniger intuitiv benutzbar: Fusion 360 oder stärker intuitiv benutzbar: SketchUp) und lud innerhalb der Software die 3D-Szene des Raumplanungsszenarios (d.h. Datei „FusionExperiment1.f3d“ für Fusion 360; Datei „SketchUpExperiment1.skp“ für SketchUp) (siehe Teilabschnitt 5.1.4.3). Die Versuchspersonen wurden nach der Aushändigung der Testaufgaben in beiden Bedingungen standardisiert mündlich vom Evaluator gebeten, die erste Aufgabe gründlich zu lesen und ihm mit einem Handzeichen zu signalisieren, dass sie mit der Aufgabenbearbeitung beginnen möchten. Im Zuge dessen wurden sie außerdem vom Evaluator standardisiert mündlich aufgefordert, diesem auch dann ein Handzeichen zu geben, wenn sie glaubten, die Testaufgabe erledigt zu haben und sie demzufolge die Aufgabenbearbeitung beenden möchten. Schließlich instruierte der Evaluator die Versuchspersonen bei der nun folgenden Bearbeitung der Testaufgabe kontinuierlich den gelernten Rhythmus mit dem USB-Fußpedal zu klopfen und die Aufgabenbearbeitung stillschweigend zu absolvieren. Für die genaue Vorgehensweise bei der Experimental-Messung wird auf die generelle Methodenbeschreibung von IntuiBeat-S im vorigen Teilabschnitt 5.1.4.3 auf Absatz „Funktionsweise des Experimental-Modus der IntuiBeat-Software und Beschreibung des Programmablaufs“ verwiesen.

Nachdem die Versuchspersonen dem Evaluator ein entsprechendes Handzeichen gegeben hatten und anschließend mit der Bearbeitung der Aufgabe begannen, startete der Evaluator daraufhin die Bildschirmaufzeichnung (inkl. Aufzeichnung der Mausbewegungen) und nahm die Rhythmusaufzeichnung bei der jeweiligen Aufgabe mithilfe der IntuiBeat-Software so vor, wie oben unter 5.1.4.3 beschrieben. Nach der Erledigung der ersten Aufgabe gaben die Versuchspersonen, wie aufgefordert, dem Evaluator ein Handzeichen und dieser stoppte daraufhin die Bildschirm- und Rhythmusaufzeichnung. Da alle Testaufgaben so ausgewählt wurden, dass ein Expertennutzer eine Aufgabe in einer Minute erledigen kann, wurde die Aufgabenbearbeitung vom Evaluator nach fünf Minuten abgebrochen. Nachdem der Evaluator die Aufzeichnungen gestoppt und diese gesichert hatte, bewertete er zunächst die Effektivität der Bearbeitung der Testaufgabe. Dieses Ergebnis trug er in eine Excel-Tabelle ein. Er instruierte die Versuchsteilnehmer anschließend die SEA-Skala und den NASA-RTLX auszufüllen, welche ihnen zu diesem Zweck auf Papier ausgehändigt wurden. Die Darbietungsreihenfolge der SEA-Skala und des NASA-RTLX wurden über alle Versuchspersonen hinweg vollständig ausbalanciert. Nachdem die Versuchspersonen die beiden Fragebögen ausgefüllt hatten, wurden sie vom Evaluator standardisiert mündlich instruiert, die Instruktion der nächsten Testaufgabe zu lesen.

Die obige Vorgehensweise zur Experimental-Messung wurde für diese und die restlichen Testaufgaben wiederholt. Nachdem die Versuchspersonen alle Testaufgaben absolviert hatten, wurden sie vom Evaluator mündlich aufgefordert die von ihnen getestete Softwareanwendung mit dem QUESI zu bewerten, der ihnen zu diesem Zweck ebenfalls auf Papier präsentiert wurde. Im Anschluss folgte ein retrospektives Interview, um Nutzungsprobleme bei der Systemnutzung zu identifizieren.

Abschlussphase (5 Minuten)

Abschließend klärte der Evaluator die Probanden standardisiert mündlich über die genaue Untersuchungsabsicht (d.h. Debriefing mit Informationen zu Studienzielen und Hypothesen) auf. Der Evaluator erkundigte sich außerdem persönlich nach dem Befinden jedes Teilnehmers und versuchte dabei mögliche Unklarheiten bezüglich des Experiments zu beseitigen. Außerdem wurde jede Versuchsperson darauf hingewiesen, dass es für den weiteren Erfolg der Untersuchung unabdingbar sei, dass gegenüber potentiellen Versuchsteilnehmern nicht über die Inhalte des Experiments gesprochen werde. Im Anschluss notierte sich der Evaluator den Namen der jeweiligen Versuchsperson auf einer Teilnehmerliste, um ihr die Versuchspersonenstunde verbuchen zu können. Um die Anonymität der Teilnehmer zu wahren, wurde diese Liste getrennt von den erhobenen Daten aufbewahrt und lag lediglich physisch vor. Des Weiteren bedankte sich der Evaluator für die Teilnahme und verabschiedete die Versuchsperson. Der Evaluator setzte am Ende des Experiments den Versuchspersonen-PC (d.h. Schließen und erneutes Öffnen von LimeSurvey zum Anlegen einer neuen Session, Schließen des getesteten CUIs) und den Analyse-PC (d.h. Schließen und erneutes Öffnen der IntuiBeat-Software) in den Ausgangszustand zurück und kontrollierte die aufgezeichneten Daten auf Vollständigkeit.

5.1.4.5 Statistische Auswertung

Die statistische Datenauswertung erfolgte mittels IBM SPSS Statistics 25 für macOS. Zunächst wurde vor der eigentlichen Datenauswertung geprüft, ob die Daten Ausreißer enthielten, die Datensätze der erhobenen Stichprobe vollständig und die Voraussetzungen der statistischen Tests erfüllt waren. Alle statistischen Tests und Analysen der Teststärke erfolgten zweiseitig.

Überprüfung der Reliabilität von IntuiBeat-S

Zunächst wurde hierzu die Testhalbierungs-Reliabilität (d.h. Überprüfung der Reliabilität) nach Guttman durch Vergleich der mittleren Rhythmusabweichungen des kurzen Rhythmusintervalls in der ersten Hälfte der Systemnutzung mit der zweiten Hälfte der Systemnutzung bestimmt, indem dieser Vergleich für die Rhythmusaufzeichnungen auf Basis der Clean-Dateien berechnet wurde. Laut Döring und Bortz (2016) variieren Reliabilitätskoeffizienten typischerweise zwischen 0 (d.h. eine vollständig unzuverlässige Messung, die nur aus zufälligen Messfehlern besteht) und 1 (d.h. eine perfekt zuverlässige Messung, ohne jegliche Beeinträchtigung durch Messfehler). Reliabilitätskoeffizienten von über oder gleich .90 wurden als hoch und Reliabilitätskoeffizienten über oder gleich .80 als akzeptabel interpretiert (siehe Bühner, 2011; Döring & Bortz, 2016).

Überprüfung der Validität von IntuiBeat-S

Im Anschluss erfolgte die Bestimmung der Konstruktvalidität anhand eines zweistufigen Vorgehens (siehe Teilabschnitt 5.1.5.3). Das Vorgehen bei der Datenauswertung wurde hier an Korbach et al. (2018) angelehnt, die die Überprüfung der wissenschaftlichen Güte

der Rhythmusmethode auf ähnliche Weise vornahmen und man so Vergleichbarkeit mit den Originalarbeiten zur Rhythmusmethode herstellen wollte. Die statistische Auswertung erfolgte dazu zum einen durch die Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H1) und zum anderen durch die Überprüfung der konvergenten und divergenten Validität von IntuiBeat-S (Hypothesen H2 und H3).

Als erster Schritt wurden die **Unterschiede in der intuitiven Gestaltung der CUIs** (Hypothese H1) multivariat im einfaktoriellen Between-Subjects Design überprüft (siehe Teilabschnitt 5.1.5.3). Die unabhängige Variable war die *unterschiedlich intuitiv benutzbare Software* mit den Ausprägungen: weniger intuitiv benutzbare Software vs. stärker intuitiv benutzbare Software. Als abhängige Variablen fungierten neben IntuiBeat-S auch die anderen konvergenten summativen Evaluationsmethoden, die das *Ausmaß an intuitiver Benutzung* auf unterschiedliche Art und Weise operationalisierten. Da die *Vorerfahrung bei der Nutzung von CUIs* und die *rhythmische Wahrnehmung*, die als Kontrollvariablen KV_Vorerfahrung und KV_Rhythmik erhoben wurden, einen ungewollten Einfluss auf den Vergleich der beiden Ausprägungen der unabhängigen Variable haben konnten, musste diese mögliche Konfundierung durch die *Vorerfahrung bei der Nutzung von CUIs* und die *rhythmische Wahrnehmung* zuerst ausgeschlossen werden. Hierzu wurden t-Tests für unabhängige Stichproben bezüglich der beiden Kontrollvariablen, sowie Pearson-Produkt-Moment-Korrelationen zwischen diesen Kontrollvariablen und den das Ausmaß an intuitiver Benutzung indizierenden konvergenten Evaluationsmethoden innerhalb beider Faktorstufen der unabhängigen Variable berechnet. Das Signifikanzniveau für alle damit verbundenen statistischen Berechnungen betrug $\alpha = .05$ und die mithilfe von G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) berechneten Effektgrößen wurden als Absolutwerte berichtet und interpretiert. Die Effektstärken für die t-Tests wurde mithilfe von G*Power (Faul et al., 2009) berechnet und als Absolutwerte interpretiert. Dabei indizierten Effektstärken d um .20 einen kleinen, um .50 einen mittleren und um .80 einen großen Effekt (J. Cohen, 1992). Für in diesem Zusammenhang bei nicht signifikanten Ergebnissen durchgeführte post-hoc Analysen der Teststärke (d.h. Poweranalysen) wurde ebenfalls auf das Programm G*Power (Faul et al., 2009) zurückgegriffen. Ein Korrelationskoeffizient $|r|$ um .10 wurde als ein schwacher, ein Korrelationskoeffizient $|r|$ um .30 als ein moderater, und ein Korrelationskoeffizient $|r|$ um .50 als hoher Zusammenhang interpretiert. Die Überprüfung der statistischen Voraussetzungen der Pearson-Produkt-Moment-Korrelationen und der t-Tests werden im folgenden Ergebnisteil berichtet (siehe Teilabschnitt 5.1.5.1).

Da hierbei keine signifikanten Unterschiede und Zusammenhänge festgestellt werden konnten (d.h. kein ungewollter Einfluss der Kontrollvariablen auf die abhängigen Variablen), wurde von einer multivariaten Kovarianzanalyse (MANCOVA) abgesehen (Döring & Bortz, 2016; Field, 2017) und sich stattdessen für eine Datenauswertung mithilfe einer multivariaten Varianzanalyse (MANOVA) ohne Berücksichtigung dieser Variablen als Kovariaten entschieden (siehe Teilabschnitt 5.1.5.3). Diese multivariate Varianzanalyse (MANOVA) wurde in Übereinstimmung mit Cramer und Bock (1966) vorgeschaltet, bevor die einzelnen Hypothesen durch mehrere univariate Varianzanalysen (ANOVAs) überprüft wurden, da diese dabei helfen kann, den Alphafehler bei darauffolgenden ANOVAs zu kontrollieren, ohne dass man das Signifikanzniveau korrigieren muss, wie es bei mehreren ANOVAS normalerweise nötig wäre. Aufgrund der geringen Stichprobengröße im Verhältnis zur hohen Anzahl an konvergenten Evaluationsmethoden würde eine Bonferroni-Korrektur oder die Korrektur mithilfe eines ähnlichen Verfahrens (z.B. Bonferroni-Holm, Hochberg) sehr zu

Lasten der Teststärke gehen (Field, 2017). Das Signifikanzniveau für alle damit verbundenen statistischen Berechnungen betrug $\alpha = .05$ und die mithilfe von IBM SPSS Statistics 25 berechneten Effektgrößen wurden als Absolutwerte berichtet und interpretiert. Bei der multivariaten Varianzanalyse wurde eine absolute Effektstärke η_p^2 um .01 als klein, eine Effektstärke η_p^2 um .06 als mittel und eine Effektstärke η_p^2 um .14 als groß interpretiert (J. Cohen, 1988). Einer signifikanten MANOVA folgten ANOVAS zur Überprüfung, um herauszufinden, ob die einzelnen konvergenten Evaluationsmethoden wirklich in der Lage sind, zwischen den Faktorstufen bezüglich des Ausmaßes intuitiver Benutzung differenzieren zu können (H1.A bis H1.F). Bei allen univariaten Varianzanalysen wurde eine absolute Effektstärke η^2 um .01 als klein, eine Effektstärke η^2 um .06 als mittel und eine Effektstärke η^2 um .14 als groß interpretiert (J. Cohen, 1988). Die Überprüfung der statistischen Voraussetzungen der MANOVA und den darauffolgenden ANOVAs werden im folgenden Ergebnisteil berichtet (siehe Teilabschnitt 5.1.5.1).

Nachdem der erste Teil des Vorgehens abgeschlossen war (Hypothese H1) und damit der Unterschied bezüglich des Ausmaßes an intuitiver Benutzung zwischen den beiden CUIs durch alle Evaluationsmethoden feststellbar war, erfolgte die **Überprüfung der konvergenten und divergenten Validität von IntuiBeat-S** als summative Evaluationsmethode intuitiver Benutzung (Hypothesen H2 und H3). Als erster Schritt wurde hier zunächst die konvergente Validität von IntuiBeat-S bestimmt (siehe Teilabschnitt 5.1.5.3), indem Pearson-Produkt-Moment-Korrelationen zwischen allen konvergenten Evaluationsmethoden, die das Ausmaß an intuitiver Benutzung abbilden, innerhalb beider Ausprägungen der unabhängigen Variable berechnet wurden. Das Signifikanzniveau betrug hier ebenfalls $\alpha = .05$. Dabei wurde geprüft, ob, wie erwartet, hohe signifikante lineare Zusammenhänge zwischen IntuiBeat-S und den konvergenten Evaluationsmethoden bestehen (H2.A bis H2.E). Als nächster Schritt wurde die divergente Validität von IntuiBeat-S bestimmt (siehe Teilabschnitt 5.1.5.3), indem ebenfalls Pearson-Produkt-Moment-Korrelationen zwischen IntuiBeat-S und allen divergenten Evaluationsmethoden, die nicht das Ausmaß an intuitiver Benutzung abbilden, innerhalb beider experimentellen Bedingungen (d.h. weniger intuitiv benutzbare Software vs. stärker intuitiv benutzbare Software) berechnet wurden. Das Signifikanzniveau betrug hier ebenfalls $\alpha = .05$. Dabei wurde geprüft, ob , wie erwartet, keine signifikanten linearen Zusammenhänge zwischen IntuiBeat-S und den anderen nicht intuitive Benutzung evaluierenden divergenten Methoden bestehen (H3.A und H3.B). Zur Überprüfung der konvergenten und divergenten Validität wurden in Anlehnung an J. Cohen (1988) ein Korrelationskoeffizient $|r|$ um .10 als ein schwacher, ein Korrelationskoeffizient $|r|$ um .30 als ein moderater, und ein Korrelationskoeffizient $|r|$ um .50 als hoher Zusammenhang interpretiert. Die Überprüfung der statistischen Voraussetzungen der Pearson-Produkt-Moment-Korrelationen werden im folgenden Ergebnisteil berichtet (siehe Teilabschnitt 5.1.5.1).

5.1.5 Ergebnisse

Im folgenden Abschnitt werden die Ergebnisse bezüglich der in Teilabschnitt 5.1.3 beschriebenen Hypothesen deskriptiv und inferenzstatistisch berichtet. Vor der eigentlichen Datenanalyse wird zunächst auf die Überprüfung der statistischen Voraussetzungen eingegangen. Dabei wurde bei allen abhängigen Variablen und Kontrollvariablen ein metrisches Skalenniveau angenommen.

5.1.5.1 Überprüfung der statistischen Voraussetzungen

Überprüfung von Ausreißern

Zur Überprüfung univariater Ausreißer wurden vor der Untersuchung der einzelnen Hypothesen modifizierte z-Werte herangezogen, da z-Werte generell ein häufig angewendetes Verfahren zur Identifikation univariater Ausreißer darstellen (Cousineau & Chartier, 2010; Shiffler, 1988). Es wurde sich speziell für die Verwendung von modifizierten z-Werten entschieden, da diese auch bei Stichproben mit geringer Größe mit hoher Zuverlässigkeit funktionieren (Garcia, 2012; Iglewicz & Hoaglin, 1993; Seo, 2006). Als univariate Ausreißer wurden in Anlehnung an Iglewicz und Hoaglin (1993) Werte identifiziert, deren absoluter modifizierter z-Wert größer als 3.5 lag. Es mussten auf diese Weise bei keiner der analysierten abhängigen Variablen und Kontrollvariablen Werte ausgeschlossen werden. An dieser Stelle ist jedoch anzumerken, dass aufgrund der geringen Varianz von KV_Vorerfahrung (siehe Tabelle 5.1) und der damit verbundenen Tatsache, dass alle Versuchspersonen nahezu keine Vorerfahrung bei der Nutzung von CUIs aufwiesen und die angegebenen Werte damit nahezu das Minimum des TFQ repräsentierten, in diesem Fall kein modifizierter z-Wert berechnet werden konnte (d.h. Dividende durch Null). Stattdessen wurde anhand eines Boxplots von KV_Vorerfahrung überprüft, inwiefern in den Daten univariate Ausreißer zu finden sind (siehe Howell, 2009). Mit diesem Verfahren konnten keine Ausreißer festgestellt werden.

Zur Überprüfung multivariater Ausreißer wurde vor Berechnung der multivariaten Varianzanalyse zur Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H1) die Mahalanobis-Distanz herangezogen, da diese ein häufig angewendetes Verfahren zur Identifikation multivariater Ausreißer darstellt (De Maesschalck, Jouan-Rimbaud, & Massart, 2000; Hadi, 1992). Anhand eines Vergleichs der Mahalanobis-Distanz mit den kritischen Werten der Chi-Quadrat-Verteilung (mit $\alpha = .001$ und bei der Gleichsetzung der Freiheitsgrade mit der Anzahl der verwendeten konvergenten sechs Evaluationsmethoden, die das Ausmaß an intuitiver Benutzung in diesem Experiment operationalisierten), wurde festgestellt, dass sich keine multivariaten Ausreißer in den Daten befanden.

Überprüfung der Voraussetzung der Normalverteilung

Die univariate Normalverteilung der abhängigen Variablen und Kontrollvariablen wurde für jede Ausprägung der unabhängigen Variable mittels Kolmogorov-Smirnov-Tests ($p \geq .05$, siehe Field, 2017) und Sichtprüfung anhand eines Q-Q-Diagramms geprüft. Dabei konnten auf Basis dieser beiden Kriterien bei den Variablen KV_Vorerfahrung, AV_CHAI

und AV_Effektivität keine Normalverteilung in beiden Gruppen im Rahmen der Überprüfung der Unterschiede bei der intuitiven Gestaltung der CUIs (Hypothese H1) und der Überprüfung der konvergenten und divergenten Validität (Hypothesen H2 und H3) festgestellt werden. Aufgrund der Tatsache, dass eine univariate Varianzanalyse bei etwa gleich großen Gruppen robust gegenüber Verletzungen der Normalverteilungsannahme ist (Blanca, Alarcón, Arnau, Bono, & Bendayan, 2017; Glass, Peckham, & Sanders, 1972; Lix, Keselman, & Keselman, 1996; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010), wurde bei der univariaten varianzanalytischen Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H1) nicht auf nonparametrische Verfahren zurückgegriffen. Da auch die zur Überprüfung der konvergenten und divergenten Validität (Hypothesen H2 und H3) eingesetzten Pearson-Produkt-Moment-Korrelationen (Edgell & Noon, 1984; Havlicek & Peterson, 1976; Liddell & Kruschke, 2018), sowie der hier zur Überprüfung der Signifikanz genutzte ungepaarte t-Test im Allgemeinen (Glass et al., 1972; Pagano, 2006; Salkind, 2010; Wilcox, 2011) robust gegenüber Verletzungen der Normalverteilungsannahme sind, wurde sich an dieser Stelle, für eine eindeutige Interpretation und Vergleichbarkeit der Ergebnisse (d.h. Mittelwerte statt Medianen) der verschiedenen Experimente, gegen Transformationen und entsprechende nonparametrische Verfahren zur Überprüfung der Hypothesen (d.h. H1, H2 und H3) entschieden.

Aufgrund der Tatsache, dass mit dem für die statistische Auswertung genutzten, IBM SPSS Statistics 25 kein multivariater Shapiro-Wilk-Test berechnet werden konnte, wurde eine Überprüfung der multivariaten Normalverteilung der abhängigen Variablen und Kontrollvariablen lediglich näherungsweise durch Überprüfung der univariaten Normalverteilung vorgenommen, so wie es eben beschrieben wurde. Da die Normalverteilungsannahme bei einigen Variablen verletzt war, erfolgte die multivariate Varianzanalyse unter Verwendung der Pillai-Spur zur Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H1), da diese nicht nur als robust gegenüber Modellverletzungen gilt, sondern auch eine große statistische Power bei kleinen Stichproben besitzt (Tabachnick et al., 2007).

Überprüfung der Voraussetzung von Linearität und Sicherstellung fehlender Multikollinearität

Die Linearität der abhängigen Variablen wurde im Rahmen der Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H1) und der Überprüfung der konvergenten und divergenten Validität (Hypothesen H2 und H3) für jede Ausprägung der unabhängigen Variable univariat anhand einer Streudiagrammmatrix überprüft und dabei keine Verletzung dieser Voraussetzung festgestellt. Die, für die multivariaten Varianzanalysen zur Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H1), außerdem erforderliche, fehlende Multikollinearität wurde innerhalb der beiden Ausprägungen der unabhängigen Variable jeweils mithilfe von Pearson-Produkt-Moment-Korrelationen zwischen den abhängigen Variablen überprüft. Aufgrund der Tatsache, dass alle Korrelationen nicht übermäßig hoch lagen ($|r| < .8$), wurde Multikollinearität ausgeschlossen (Berry, Feldman, & Stanley Feldman, 1985; Rockwell, 1975). Es wurden an dieser Stelle trotzdem Pearson-Produkt-Moment-Korrelationen berechnet, obwohl die Normalverteilungsannahme einiger Variablen (siehe „Überprüfung der Voraussetzung der

Normalverteilung“ des Teilabschnitts 5.1.5.1) verletzt war, da diese bei etwa gleich großen Gruppen gegenüber dieser Verletzung relativ robust reagieren (Edgell & Noon, 1984; Havlicek & Peterson, 1976; Liddell & Kruschke, 2018).

Überprüfung der Voraussetzung der Homoskedastizität und der Homogenität der Varianz-Kovarianzen

Zur Überprüfung der Homoskedastizität zwischen den Ausprägungen der unabhängigen Variable kamen Levene-Tests zum Einsatz (siehe Field, 2017), welche im Rahmen der univariaten varianzanalytischen Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H1) berechnet und bei allen abhängigen Variablen und Kontrollvariablen ($p \geq .05$, siehe Field, 2017) außer bei AV_QUESI und AV_Effektivität Varianzhomogenität bestätigen konnten. Da dementsprechend für diese beiden Variablen keine homogenen Varianzen vorlagen, wurde sich für die Berechnung von Welch-ANOVAs entschieden, da diese im Gegensatz zu klassischen ANOVAs keine Varianzhomogenität voraussetzen und eine hohe Teststärke bei gleichzeitig geringer Wahrscheinlichkeit eines Typ 1 Fehlers aufweisen (Moder, 2007, 2010). Aufgrund der Tatsache, dass sich die Effektstärke bei einer Welch-ANOVA nicht ohne Präzisionsverlust als η^2 angeben lässt (Olejnik & Algina, 2003), wurde sich für die Angabe der Effektstärke zusätzlich in Form von ω^2 entschieden, da diese eine konservative Einschätzung der Effektstärke bei Verletzung der Varianzhomogenität darstellt (Skidmore & Thompson, 2013) und sich als Schätzgröße für die durch die Variable aufgeklärte Varianz auch auf die gleiche Weise wie η^2 interpretieren lässt (d.h. ω^2 fällt nur kleiner aus, da mit η^2 der Effekt häufig überschätzt wird, siehe C. Albers & Lakens, 2018; Okada, 2013) (siehe J. Cohen, 1988). Zur Überprüfung der Voraussetzung der Homogenität der Varianz-Kovarianzen wurde ein Boxscher M-Test im Rahmen der Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H1) berechnet, der nicht signifikant war ($p \geq .05$, siehe Field, 2017) und dementsprechend die Varianz-Kovarianzen als homogen anzusehen sind.

Überprüfung der Voraussetzung der Unabhängigkeit der Fehlerkomponenten

Laut Bortz und Schuster (2011) kann diese Voraussetzung als gegeben angesehen werden, wenn die Versuchspersonen den einzelnen Ausprägungen der unabhängigen Variable randomisiert zugewiesen wurden, was im hier berichteten ersten Experiment der Fall war.

5.1.5.2 Überprüfung der Reliabilität von IntuiBeat-S

Wie erwartet, zeigte die Analyse der Testhalbierungs-Reliabilität der mittleren Rhythmusabweichungen über alle Versuchspersonen hinweg Guttman Testhalbierungs-Koeffizienten von $r_{Fusion360} = .983$ (d.h. weniger intuitiv benutzbare Software) und $r_{SketchUp} = .980$ (d.h. stärker intuitiv benutzbare Software), was laut Döring und Bortz (2016) als eine hohe Reliabilität von IntuiBeat-S interpretiert werden kann.

5.1.5.3 Überprüfung der Validität von IntuiBeat-S

Tabelle 5.1. *Deskriptive Daten und Teststatistiken des ersten Experiments.*

	Intuitiv ↓ Fusion 360		Intuitiv ↑ SketchUp		Teststatistik
	M	*SD*	*M*	*SD*	
AV_IntuiBeat-S [ms]	107.04	35.43	77.43	26.13	$F(1,30) = 7.24$ $p = .012$ $\eta^2 = .19$
AV_CHAI [s]	.14	.14	.30	.23	$F(1,30) = 5.42$ $p = .027$ $\eta^2 = .15$
AV_Effektivität [%]	62.50	29.50	52.08	45.49	$F(1,25.72) = .59$ $p = .449$ $\omega^2 = .01$ oder $\eta^2 = .02$
AV_QUESI [1-5]	1.92	.77	2.62	1.07	$F(1,27.26) = 4.48$ $p = .043$ $\omega^2 = .10$ oder $\eta^2 = .13$
AV_NASA [0-100]	56.89	18.14	44.23	11.74	$F(1,30) = 5.50$ $p = .026$ $\eta^2 = .16$
AV_SEA [0-220]	112.42	32.56	88.95	22.67	$F(1,30) = 5.60$ $p = .03$ $\eta^2 = .16$
AV_PhysischeEffizienz [%]	49.17	16.80	31.85	16.66	-
AV_ZeitlicheEffizienz [s]	99.25	61.82	64.56	56.05	-
KV_Rhythmik [0-8.67]	4.59	1.63	4.94	1.47	$t(30) = -.63$ $p = .536$ d = .22
KV_Vorerfahrung [0-6]	.03	.10	.06	.14	$t(30) = .87$ $p = .392$ d = .31

- : Es erfolgte keine inferenzstatistische Auswertung, da es sich hierbei um divergente Quasi-Außenkriterien handelte, die keine Unterschiede in der intuitiven Gestaltung der CUIs zeigen sollten.

Überprüfung einer möglichen Konfundierung durch die Vorerfahrung bei der Nutzung von CUIs und durch die rhythmische Wahrnehmung

Die Vorerfahrung bei der Nutzung von CUIs und die rhythmische Wahrnehmung bei der weniger intuitiv benutzbaren Software unterschied sich nicht signifikant ($p > .05$) von der Vorerfahrung bei der Nutzung von CUIs bei der stärker intuitiv benutzbaren Software (siehe Tabelle 5.1). Laut J. Cohen (1988) können beide Effekte als eher klein interpretiert werden ($d \leq .5$). Eine konservative post-hoc Analyse der Teststärke bezüglich KV_Vorerfahrung und KV_Rhythmik mithilfe von G*Power (Faul et al., 2009) mit $df = 30$ und einer angenommenen geringen Effektstärke ($d_{KV\_Vorerfahrung} = .31$ bzw. $d_{KV\_Rhythmik} = .22$) ergab lediglich eine geringe Teststärke (1 - $\beta_{KV\_Vorerfahrung} = .13$ bzw. 1 - $\beta_{KV\_Rhythmik} = .09$) im Zuge der Auswertung der Kontrollvariablen. Um jedoch eine ausreichend große Power (1 - $\beta \geq .80$) beim festgestellten Effekt erzielen zu können (siehe J. Cohen, 1988), wären pro Gruppe 168 bzw. 326 Versuchspersonen nötig gewesen, was aufgrund des straffen Zeitplans im Anwenderprojekt, des dortigen Fokus auf qualitative Ergebnisse, des Verständnisses der Meta-Evaluation von IntuiBeat-S als Nebenprodukt und der personellen Einschränkungen nicht möglich gewesen wäre. Auch bei Annahme einer großen Effektstärke (d = .8), die aufgrund der geringen Anforderungen an CUI-Kenntnisse bei der Stichprobe auch nicht zu erwarten wäre, hätten immerhin noch 26 Versuchspersonen pro Gruppe getestet werden müssen, was im Hinblick auf die Einschränkungen durch das Anwenderprojekt nicht notwendig erschien.

Da die erhobene Vorerfahrung bei der Nutzung von CUIs und die rhythmische Wahrnehmung dennoch einen ungewollten Einfluss auf den Vergleich der beiden Ausprägungen der unabhängigen Variable haben könnte, wurde für jede der beiden Kontrollvariablen mithilfe von Pearson-Produkt-Moment-Korrelationen sichergestellt, dass kein linearer Zusammenhang zwischen der jeweiligen Kontrollvariable und den das Ausmaß intuitiver Benutzung abbildenden Evaluationsmethoden innerhalb der beiden Ausprägungen der unabhängigen Variable besteht. Die beiden Kontrollvariablen wurden dementsprechend nicht als Kovariaten in den folgenden Varianzanalysen berücksichtigt (siehe Döring & Bortz, 2016; Field, 2017).

Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs

Eine multivariate Varianzanalyse zeigte erwartungsgemäß einen signifikanten Effekt der unabhängigen Variable *unterschiedlich intuitiv benutzbaren Software* (d.h. weniger intuitiv benutzbare Software vs. stärker intuitiv benutzbare Software) auf das durch die Evaluationsmethoden erfasste Ausmaß an intuitiver Benutzung, $F(6,25) = 3.63$, $p < .05$, Pillai-Spur = .47, $\eta_p^2 = .47$. Laut J. Cohen (1988) handelt es sich es hierbei um einen großen Effekt ($\eta_p^2 \geq .14$).

Die weitere univariate varianzanalytische Auswertung der einzelnen als Quasi-Außenkriterien fungierenden Evaluationsmethoden ergab für alle konvergenten Quasi-Außenkriterien (H1.A, H1.B, H1.C, H1.D und H1.F) bis auf die Messung der Effektivität (AV_Effektivität; H1.E) statistisch signifikante Effekte (siehe Tabelle 5.1), weswegen die Manipulation (d.h. CUIs unterscheiden sich in ihrer intuitiven Benutzung) als erfolgreich interpretiert werden konnte und auch die eingesetzten Evaluationsmethoden (bis auf AV_Ef-

fektivität) empirisch als konvergente Quasi-Außenkriterien für die Meta-Evaluation von IntuiBeat-S bestätigt werden konnten (d.h. alle AVs bis auf AV_Effektivität konnten den Unterschied zwischen den unterschiedlich intuitiv benutzbaren CUIs identifizieren). Eine konservative post-hoc Analyse der Teststärke mithilfe von G*Power mit $df = 1$, sowie der Annahme einer Nullkorrelation innerhalb der unabhängigen Variable ergab bei einer angenommenen geringen Effektstärke ($\eta^2_{AV\_Effektivität} = .02$) eine zu geringe Teststärke (1 - $\beta_{AV\_Effektivität} = .12$) im Zuge der Auswertung der abhängigen Variable AV_Effektivität.

Obwohl die AV_Effektivität auf Basis des nicht signifikanten Ergebnisses nicht empirisch als konvergentes Quasi-Außenkriterium bestätigt werden konnte (siehe Tabelle 5.1), wurde diese abhängige Variable dennoch als konvergentes Quasi-Außenkriterium berücksichtigt. Es wurde sich dafür entschieden, da die AV_Effektivität in beiden Gruppen gleich grob (d.h. eine Aufgabe wurde nur dichotom als erfolgreich beendet oder nicht bewertet) operationalisiert wurde und laut Döring und Bortz (2016) die interne Validität dadurch nicht reduziert wurde. Obwohl wegen der Operationalisierung der AV_Effektivität nur große Unterschiede im Ausmaß intuitiver Benutzung erkennbar sein sollten (dafür spricht auch die geringe Effektstärke), sollte trotzdem ein Zusammenhang zwischen IntuiBeat-S und der Effektivität erkennbar sein. Laut J. Cohen (1988) handelte es sich bei den Effekten, die bei den konvergenten Quasi-Außenkriterien beobachtet werden konnten, um große ($\eta^2 \geq .14$) und um mittlere ($\eta^2 \geq .06$) Effekte, lediglich bezüglich der AV_Effektivität konnte ein kleiner Effekt festgestellt werden ($\eta^2 = .02$). Die univariate varianzanalytische Auswertung der beiden Ausprägungen der unabhängigen Variable bezüglich IntuiBeat-S ergab erwartungsgemäß ebenfalls einen signifikanten Effekt (siehe Tabelle 5.1), der laut J. Cohen (1988) als groß ($\eta^2 \geq .14$) interpretiert werden kann. Die deskriptiven Daten der multivariaten und univariaten Varianzanalysen sind der Tabelle 5.1 zu entnehmen.

Überprüfung der konvergenten Validität von IntuiBeat-S

Tabelle 5.2. *Pearson-Produkt-Moment-Korrelationen zwischen summativen Evaluationsmethoden für intuitive Benutzung (d.h. konvergente Quasi-Außenkriterien) und IntuiBeat-S zur Überprüfung der konvergenten Validität (Hypothese H2) im Zuge des ersten Experiments.*

	AV_IntuiBeat-S: Intuitiv ↓ Fusion 360	AV_IntuiBeat-S: Intuitiv ↑ SketchUp
AV_CHAI	$r(16) = -.59$ $p = .017$	$r(16) = -55$ $p = .028$
AV_Effektivität	$r(16) = -.60$ $p = .014$	$r(16) = -.54$ $p = .033$
AV_QUESI	$r(16) = -.55$ $p = .029$	$r(16) = -.66$ $p = .008$
AV_NASA	$r(16) = .58$ $p = .019$	$r(16) = .56$ $p = .025$
AV_SEA	$r(16) = .64$ $p = .008$	$r(16) = .55$ $p = .028$

Wie erwartet, zeigten die Pearson-Produkt-Moment-Korrelationen (siehe Tabelle 5.2) zwischen den durch IntuiBeat-S erhobenen mittleren Rhythmusabweichungen und den Testwerten der anderen, für die Meta-Evaluation fungierenden, konvergenten Quasi-Außenkriterien, im Rahmen der Überprüfung der konvergenten Validität (Hypothese H2), bei beiden Ausprägungen der unabhängigen Variable signifikante Korrelationskoeffizienten von $|r| > .50$ ($p < .05$), weswegen hier laut J. Cohen (1988) von hohen linearen Zusammenhängen ($|r| \geq .5$) zwischen den konvergenten Evaluationsmethoden und IntuiBeat-S (H2.A bis H2.E) gesprochen werden kann.

Überprüfung der divergenten Validität von IntuiBeat-S

Wie erwartet, zeigten die Pearson-Produkt-Moment-Korrelationen (siehe Tabelle 5.3) zwischen den durch IntuiBeat-S erhobenen mittleren Rhythmusabweichungen und den Testwerten der, nicht intuitive Benutzung evaluierenden, divergenten Methoden (d.h. physische Effizienz bei der Systemnutzung und zeitliche Effizienz bei der motorischen Handlungsdurchführung), im Rahmen der Überprüfung der divergenten Validität (Hypothese H3), bei beiden Ausprägungen der unabhängigen Variable allesamt (H3.A und H3.B) nicht signifikante lineare Zusammenhänge mit IntuiBeat-S ($p \geq .05$).

Tabelle 5.3. *Pearson-Produkt-Moment-Korrelationen zwischen summativen Evaluationsmethoden, die nicht intuitive Benutzung messen (d.h. divergente Quasi-Außenkriterien) und IntuiBeat-S zur Überprüfung der divergenten Validität (Hypothese H3) im Zuge des ersten Experiments.*

	AV_IntuiBeat-S: Intuitiv ↓ Fusion 360	AV_IntuiBeat-S: Intuitiv ↑ SketchUp
AV_PhysischeEffizienz	$r(16) = .20$ $p = .950$	$r(16) = .35$ $p = .184$
AV_ZeitlicheEffizienz	$r(16) = .38$ $p = .149$	$r(16) = .04$ $p = .886$

5.1.6 Diskussion

Im vorliegenden ersten Experiment wurde die wissenschaftliche Güte von IntuiBeat-S bezüglich des Gütekriteriums *Validität* im Vergleich mit weiteren, als konvergente und divergente Quasi-Außenkriterien fungierenden, summativen Evaluationsmethoden bei der Nutzung einer weniger intuitiv benutzbaren (Fusion 360, Autodesk, 2017a) und einer stärker intuitiv benutzbaren (SketchUp, Trimble Navigation Ltd., 2016) Software empirisch geprüft.

Anhand der statistischen Tests wurde zunächst überprüft, inwiefern die Unterschiede in der intuitiven Gestaltung der CUIs mithilfe der erhobenen konvergenten Quasi-Außenkriterien und IntuiBeat-S festgestellt werden können (Hypothese H1). Abschließend wurde untersucht, inwiefern IntuiBeat-S konvergente Validität (Hypothese H2) und divergente Validität (Hypothese H3) unter Berücksichtigung konvergenter und divergenter Quasi-Außenkriterien attestiert werden kann. Im folgenden Verlauf werden die Ergebnisse bezüglich der

Reliabilität von IntuiBeat-S, sowie die einzelnen Hypothesen bezüglich der Validität unter Berücksichtigung der festgestellten Ergebnisse diskutiert, nachdem auf eine mögliche Konfundierung durch die *Vorerfahrung bei der Nutzung von CUIs* und durch die *rhythmische Wahrnehmung* eingegangen wurde.

5.1.6.1 Überprüfung der Reliabilität

Mit einer in Form eines Guttman Reliabilitätskoeffizienten berechneten Testhalbierungsreliabilität von $r_{Fusion360} = .983$ (d.h. weniger intuitiv benutzbare Software) und $r_{SketchUp} = .980$ (d.h. stärker intuitiv benutzbare Software), die laut Döring und Bortz (2016) als hoch zu interpretieren ist, und sich in einem ähnlichen Wertebereich wie die Ursprungsstudie zur Rhythmusmethode ($r_{erstesExperiment} = .96$ und $r_{zweitesExperiment} = .72$, siehe Park & Brünken, 2015) und deren Folgestudie ($r = .96$, siehe Korbach et al., 2018) befindet, kann IntuiBeat-S als summative Evaluationsmethode für intuitive Benutzung Reliabilität attestiert und damit das entsprechende Gütekriterium als gegeben angesehen werden.

5.1.6.2 Überprüfung der Validität

Überprüfung einer möglichen Konfundierung durch die Vorerfahrung bei der Nutzung von CUIs und die rhythmische Wahrnehmung

Die Vorerfahrung bei der Nutzung von CUIs und die rhythmische Wahrnehmung unterschied sich erwartungsgemäß nicht signifikant zwischen den beiden Ausprägungen der unabhängigen Variable *unterschiedlich intuitiv benutzbare Software*. Jedoch war die Teststärke aufgrund der geringen beobachteten Effekte und der kleinen Stichprobe gering. Eine mögliche Erklärung könnte bereits in der Auswahl der Stichprobe bestehen, da bei den Studierenden keine großen Unterschiede bezüglich der Vorerfahrung und der rhythmischen Wahrnehmung existieren. Dementsprechend ist die praktische Bedeutsamkeit, der in diesem Zusammenhang ermittelten, Effekte bzw. der Erkenntnisgewinn eher als unbedeutend einzustufen (siehe Döring & Bortz, 2016), was sich auch in den deskriptiven Unterschieden erkennen lässt (siehe Tabelle 5.1). Da darüber hinaus bei beiden Kontrollvariablen keine signifikanten Korrelationen mit den, das Ausmaß an intuitiver Benutzung erfassenden, konvergenten Evaluationsmethoden innerhalb beider Ausprägungen der unabhängigen Variable festgestellt werden konnten, lag mit hoher Wahrscheinlichkeit keine Konfundierung durch eine unterschiedliche Vorerfahrung bei der Nutzung von CUIs und/oder durch eine unterschiedliche rhythmische Wahrnehmung vor.

Da die im Rahmen des ersten Experiments vorgenommene Operationalisierung der Vorerfahrung (KV_Vorerfahrung) lediglich die Häufigkeits- und nicht die Funktionsdimension (d.h. Funktionsumfang) des TFQ berücksichtigte und es im Gegensatz zur rhythmischen Wahrnehmung bei anderen CUIs weiterhin zu Konfundierungen der Ergebnisse durch unterschiedliche Vorerfahrung kommen kann, soll in Folgeexperimenten diese Kontrollvariable weiterhin berücksichtigt und anhand beider Dimensionen des TFQ operationalisiert werden.

Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs

Wie erwartet konnten die als konvergente Quasi-Außenkriterien fungierenden Evaluationsmethoden QUESI, SEA, NASA-RTLX und CHAI signifikant ein höheres Ausmaß an intuitiver Benutzung bezüglich der stärker intuitiv benutzbaren Software im Vergleich zur weniger intuitiv benutzbaren Software feststellen (siehe Tabelle 5.1). Es ließen sich hierbei hohe und mittlere Effekte beobachten, wodurch sich diese Evaluationsmethoden auch empirisch als konvergente Quasi-Außenkriterien für die Meta-Evaluation von IntuiBeat-S qualifizieren und die in diesem Zusammenhang aufgestellten Hypothesen (H1.A, H1.B, H1.C, H1.D und H1.F) als bestätigt angesehen werden können (siehe Tabelle 5.4).

Entgegen der Erwartungen konnte jedoch durch die Messung der Effektivität kein signifikantes Ergebnis bezüglich dieses gerichteten prognostizierten Unterschiedes ermittelt, und die damit verbundene Hypothese H1.E nicht bestätigt werden (siehe Tabelle 5.4). Eine mögliche Erklärung hierfür kann in der Operationalisierung der abhängigen Variable AV_Effektivität liegen. Eine Aufgabe wurde hierbei nur dichotom dahingehend bewertet, ob sie erfolgreich absolviert wurde oder nicht. Feinere Unterschiede in der Aufgabenbewältigung, so wie sie auf Basis der deskriptiven Unterschiede der anderen konvergenten Quasi-Außenkriterien in diesem Experiment zu erwarten wären (siehe Tabelle 5.1), sind anhand dieser groben Operationalisierung kaum erkennbar, wofür auch die bei diesem Maß kleine Effektstärke im Vergleich zu den höheren Effektstärken der anderen als konvergente Quasi-Außenkriterien fungierenden Evaluationsmethoden spricht. Dementsprechend ist in Folgestudien eine andere Operationalisierung für die AV_Effektivität zu berücksichtigen, mit der auch kleinere Unterschiede in der intuitiven Gestaltung von Softwarepaketen bezüglich des Ausmaßes an intuitiver Benutzung feststellbar sind.

Überprüfung der konvergenten Validität von IntuiBeat-S

Wie erwartet zeigten sich hohe signifikante lineare Zusammenhänge (d.h. konvergente Validitätskoeffizienten) zwischen IntuiBeat-S und allen, als Quasi-Außenkriterien fungierenden, konvergenten Evaluationsmethoden, weswegen alle in diesem Zusammenhang aufgestellten Hypothesen bestätigt werden konnten (H2.A bis H2.E, siehe Tabelle 5.4).

Überprüfung der divergenten Validität von IntuiBeat-S

Wie erwartet zeigten sich nicht signifikant lineare Zusammenhänge (d.h. divergente Validitätskoeffzienten) zwischen IntuiBeat-S und den, nicht intuitive Benutzung evaluierenden, divergenten Methoden, weswegen alle in diesem Zusammenhang aufgestellten Hypothesen bestätigt werden konnten (H3.A und H3.B, siehe Tabelle 5.4). Dennoch besteht in der Operationalisierung der physischen Effizienz als Anzahl der Klicks bei der Systemnutzung die Gefahr, dass dieses Maß mit Evaluationsmethoden, die die intuitive Benutzung abbilden, bei einer stärker intuitiv benutzbaren Software ungewollt korrelieren könnte. Diese Gefahr konnte mithilfe von Pearson-Produkt-Moment-Korrelationen ($p > .05$) jedoch ausgeschlossen werden.

5.1.7 Schlussfolgerung

Zusammenfassend kann festgehalten werden, dass die wissenschaftliche Güte von IntuiBeat-S als summative Evaluationsmethode für intuitive Benutzung hinsichtlich der Gütekriterien *Reliabilität* und *Validität* empirisch bestätigt werden konnte (siehe Tabelle 5.4). Konfundierungen durch Unterschiede in der *Vorerfahrung bei der Nutzung von CUIs* und durch *Unterschiede in der rhythmischen Wahrnehmung* können mit hoher Wahrscheinlichkeit ausgeschlossen werden. Die Effekt- und Teststärken lagen dabei überwiegend im oberen Bereich. Lediglich bezüglich der abhängigen Variable AV_Effektivität konnte eine kleine Effektstärke, eine geringe Teststärke und ein nicht signifikanter Effekt festgestellt werden (siehe Tabelle 5.4). Neben der in diesem Zusammenhang bereits angesprochenen zu groben Operationalisierung der Effektivität, durch die nur die erfolgreich abgeschlossenen Aufgaben gewertet und damit keine detailliertere Analyse der Effektivität des Systems vorgenommen wurde, lässt sich noch eine weitere Einschränkung des ersten Experiments erkennen. Da alle konvergenten Quasi-Außenkriterien auf Basis der Daten eines Nutzertests mit zusätzlicher paralleler Rhythmusaufgabe erhoben wurden, könnte es auch zu Konfundierungen dieser Maße aufgrund der Rhythmuszweitaufgabe und einer damit verbundenen Intrusion gekommen sein.

Tabelle 5.4. *Übersicht der mithilfe des ersten Experiments bestätigten Hypothesen im Zuge der Meta-Evaluation von IntuiBeat-S.*

Hypothese	Experiment 1
(H1) Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs:	
- (A) IntuiBeat-S	✓
- (B) SEA	✓
- (C) NASA-RTLX	✓
- (D) QUESI	✓
- (E) Effektivität	✗
- (F) CHAI	✓
(H2) Überprüfung der konvergenten *Validität:*	
- (A) SEA	✓
- (B) NASA-RTLX	✓
- (C) QUESI	✓
- (D) Effektivität	✓
- (E) CHAI	✓
	✓
(H3) Überprüfung der divergenten Validität:	
- (A) Physische Effizienz	✓
- (B) Zeitliche Effizienz	✓

Obwohl basierend auf den theoretischen Überlegungen in Zusammenhang mit dem Modell multipler Ressourcen von Wickens (2008) die Wahrscheinlichkeit gering ist, dass eine auf Inhibitionsprozessen basierende und modalitätsunabhängige Zweitaufgabe, wie sie bei IntuiBeat-S eingesetzt wird, intrusiv ist (siehe Teilabschnitt 3.6.2) und sich das auch unterschiedlich auf die beiden Ausprägungen der unabhängigen Variable auswirken könnte

(d.h. bei der weniger intuitiv benutzbaren Software sollte die Intrusion durch die Zweitaufgabe aufgrund der insgesamt höheren mentalen Beanspruchung eine höhere Auswirkung haben), wird in einem Folgeexperiment empirisch untersucht, ob eine solche Zweitaufgabe die als Quasi-Außenkriterien genutzten konvergenten Evaluationsmethoden ungewollt beeinflusst. Wie bereits in Teilabschnitt 4.1 angesprochen, konnte sowohl das Ursprungsexperiment (Park & Brünken, 2015) als auch das Folgeexperiment (Korbach et al., 2018) nahezu keine signifikanten Zusammenhänge zwischen der Rhythmusmethode und den mentale Beanspruchung messenden Quasi-Außenkriterien (d.h. subjektive Evaluationsmethoden: Fragebogen, wobei hier zumindest das die Aufgabenschwierigkeit erfassende Item signifikant wurde; objektive Evaluationsmethoden: diverse Maße für Augenbewegungen, die allesamt nicht signifikant wurden) und auch keine signifikanten Korrelationen zwischen den subjektiven und objektiven Maßen feststellen. Dies könnte auch auf eine Verzerrung durch die zusätzliche Zweitaufgabe und die Verwendung des damit zusammenhängenden Datenmaterials hinweisen. Eine konzeptuelle Replikation des ersten Experiments mit (1) zwei anderen unterschiedlich intuitiv benutzbaren CUIs, (2) der Berücksichtigung der Art des als Datengrundlage genutzten Nutzertests (d.h. Nutzertest ohne Rhythmusaufgabe vs. Nutzertest mit Rhythmusaufgabe) und einer (3) feineren Operationalisierung der AV_Effektivität sollte alle offenen Fragen klären können, die damit in Zusammenhang stehen.

5.2 Experiment 2

Das zweite Experiment verfolgte das Ziel, zu überprüfen, inwiefern IntuiBeat-S auch wissenschaftliche Güte bezüglich des formalen Hauptgütekriteriums *Validität* im Vergleich mit weiteren, als konvergente, sowie divergente Quasi-Außenkriterien fungierenden, summativen Evaluationsmethoden bei der Nutzung eines weniger intuitiv benutzbaren (Fusion 360, Autodesk, 2017a) und eines stärker intuitiv benutzbaren (Affinity Designer, Serif Inc., 2017) CUI aufweist, wenn die konvergenten und divergenten Quasi-Außenkriterien nicht auf Basis eines für die Anwendung von IntuiBeat-S modifizierten Nutzertests mit Rhythmusaufgabe erhoben werden, sondern auf Basis eines gewöhnlichen Nutzertests ohne zusätzliche Rhythmusnebenaufgabe. Obwohl, wie im ersten Experiment (siehe Abschnitt 5.1) beschrieben, auf Inhibitionsprozessen basierende Zweitaufgaben unter Berücksichtigung des Modells multipler Ressourcen (Wickens, 2008) theoretisch die nötige Interferenz erzeugen können, um Schwankungen in der mentalen Beanspruchung bei gleichzeitig minimaler Intrusion sichtbar zu machen, und die Verwendung eines, für die Anwendung von IntuiBeat-S, modifizierten Nutzertests mit Rhythmusaufgabe keinen Einfluss auf die eigentliche Aufgabenbewältigung haben sollte, soll diese Vermutung im Rahmen des zweiten Experiments empirisch untersucht und eine mögliche Intrusion beim Einsatz von IntuiBeat-S auf dieses Weise ausgeschlossen werden. Gleichzeitig soll die zeitliche Anwendungseffizienz von IntuiBeat-S bewertet werden.

Da ansonsten keine Änderungen an IntuiBeat-S vorgenommen wurden, entfiel auch eine erneute Untersuchung der formalen wissenschaftlichen Gütekriterien *Objektivität* und *Reliabilität*. Neben der wissenschaftlichen Güte von IntuiBeat-S als summative Evaluationsmethode intuitiver Benutzung, sollte im Rahmen des zweiten Experiments außerdem nachgewiesen werden, inwiefern sich IntuiBeat-S, welche als Basis einen Nutzertest mit

Rhythmusaufgabe benötigt, von der CHAI-Methode, die als Basis einen Nutzertest ohne Rhythmusaufgabe benötigt, bezüglich der zeitlichen Anwendungseffizienz unterscheidet. Auf diese Weise wurde sichergestellt, dass IntuiBeat-S dementsprechend auch ein wichtiger Aspekt praktischer Güte aus Anwenderprojektsicht attestiert werden kann. Es wurde somit die erste Forschungsfrage und der summative Aspekt der dritten Forschungsfrage dieser Arbeit durch das zweite Experiment betrachtet.

5.2.1 Überprüfung der Validität von IntuiBeat-S

Aufgrund der Tatsache, dass im Vergleich zum im Abschnitt 5.1 beschriebenen Vorgehen zur Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs und zur Überprüfung der konvergenten und divergenten Validität zusätzlich noch die Art des als Basis genutzten Nutzertests berücksichtigt werden sollte, wurden die Hypothesen des ersten Experiments (siehe Teilabschnitt 5.1.3) für das zweite Experiment dementsprechend erweitert und modifiziert. Demnach kann IntuiBeat-S nur Konstruktvalidität attestiert werden, wenn die Art des als Basis für die Validierung genutzten Nutzertests unerheblich ist, da auf diese Weise sichergestellt ist, dass die zusätzlich Rhythmusaufgabe die Systemnutzung in keinster Weise verzerrt und IntuiBeat-S damit die theoretisch versprochene Interferenz bei gleichzeitig minimaler Intrusion besitzt.

5.2.1.1 Hypothesen

Unter Berücksichtigung obiger Überlegungen lassen sich die Hypothesen des ersten Experiments zur Überprüfung der Konstruktvalidität von IntuiBeat-S im Rahmen des zweiten Experiments folgendermaßen anpassen:

H1 (Überprüfung der Intrusion durch die Art des Nutzertests)

- H1.A (Haupteffekt des Innersubjektfaktors *unterschiedlich intuitiv benutzbarer Software*) Bei der stärker intuitiv benutzbaren Software ist das mit den verschiedenen als, konvergente Quasi-Außenkriterien fungierenden Evaluationsmethoden erfasste, Ausmaß an intuitiver Benutzung höher als bei der weniger intuitiv benutzbaren Software.

- H1.B (Kein Haupteffekt des Zwischensubjektfaktors *Art des Nutzertests*) Beim Nutzertest mit Rhythmusaufgabe und beim Nutzertest ohne Rhythmusaufgabe zeigen sich bezüglich des durch die verschiedenen als, konvergente Quasi-Außenkriterien fungierenden Evaluationsmethoden erfassten, Ausmaßes an intuitiver Benutzung keine Unterschiede.

- H1.C (Kein Interaktionseffekt des Zwischensubjektfakors *Art des Nutzertests* und des Innersubjektfaktors *unterschiedlich intuitiv benutzbarer Software*) Beim Nutzertest mit Rhythmusaufgabe und beim Nutzertest ohne Rhythmusaufgabe zeigen sich bezüglich des, durch die verschiedenen als konvergente Quasi-Außenkriterien fungierenden Evaluationsmethoden erfassten, Ausmaßes an intuitiver Benutzung auch innerhalb der unterschiedlich intuitiv benutzbaren Software keine Unterschiede.

H2 (Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs) Bei der stärker intuitiv benutzbaren Software liegt das, mit den verschiedenen konvergenten Evaluationsmethoden erfasste, Ausmaß an intuitiver Benutzung bei der Verwendung eines Nutzertests mit Rhythmusaufgabe (d.h. bedingter Haupteffekt des Innersubjektsfaktors *unterschiedlich intuitive benutzbare Software* unter der Bedingung *Nutzertest mit Rhythmusaufgabe* des Zwischensubjektfaktors *Art des Nutzertests*) höher als bei der weniger intuitiv benutzbaren Software.

- **H2.A (IntuiBeat-S)** Bei der stärker intuitiv benutzbaren Software sind die Rhythmusabweichungen mit IntuiBeat-S bei der Verwendung eines Nutzertests mit Rhythmusaufgabe geringer als bei der weniger intuitiv benutzbaren Software.

- **H2.B (SEA)** Bei der stärker intuitiv benutzbaren Software sind die Ratings der SEA-Skala bei der Verwendung eines Nutzertests mit Rhythmusaufgabe geringer als bei der weniger intuitiv benutzbaren Software.

- **H2.C (NASA-RTLX)** Bei der stärker intuitiv benutzbaren Software sind die Gesamtratings des NASA-RTLX bei der Verwendung eines Nutzertests mit Rhythmusaufgabe geringer als bei der weniger intuitiv benutzbaren Software.

- **H2.D (QUESI)** Bei der stärker intuitiv benutzbaren Software sind die Gesamtratings des QUESI bei der Verwendung eines Nutzertests mit Rhythmusaufgabe höher als bei der weniger intuitiv benutzbaren Software.

- **H2.E (Effektivität)** Bei der stärker intuitiv benutzbaren Software sind die Anteile korrekt abgeschlossener Aufgaben an der Gesamtzahl abgeschlossener Aufgaben bei der Verwendung eines Nutzertests mit Rhythmusaufgabe höher als bei der weniger intuitiv benutzbaren Software.

- **H2.F (CHAI)** Bei der stärker intuitiv benutzbaren Software sind die Anteile aller intuitiven Mausklicks an der Gesamtzahl der getätigten Mausklicks bei der Systemnutzung bei der Verwendung eines Nutzertests mit Rhythmusaufgabe höher als bei der weniger intuitiv benutzbaren Software.

H3 (Überprüfung der konvergenten Validität) Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und den, mit den verschiedenen konvergenten Evaluationsmethoden erfassten, Ausmaßen an intuitiver Benutzung bestehen bei der Verwendung eines Nutzertests mit Rhythmusaufgabe hohe lineare Zusammenhänge.

- **H3.A (SEA)** Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und den Ratings der SEA-Skala besteht bei der Verwendung eines Nutzertests mit Rhythmusaufgabe ein positiver, hoher, linearer Zusammenhang.

- **H3.B (NASA-RTLX)** Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und den Gesamtratings des NASA-TLX besteht bei der Verwendung eines Nutzertests mit Rhythmusaufgabe ein positiver, hoher, linearer Zusammenhang.

- **H3.C (QUESI)** Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und den Gesamtratings des QUESI besteht bei der Verwendung eines Nutzertests mit Rhythmusaufgabe ein negativer, hoher, linearer Zusammenhang.

- **H3.D (Effektivität)** Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und dem Anteil korrekt abgeschlossener Aufgaben an der Gesamtzahl abgeschlossener Aufgaben bei der Systemnutzung besteht bei der Verwendung eines

Nutzertests mit Rhythmusaufgabe ein negativer, hoher, linearer Zusammenhang.

- **H3.E (CHAI)** Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und dem Anteil aller intuitiven Mausklicks an der Gesamtzahl der getätigten Mausklicks bei der Systemnutzung besteht bei der Verwendung eines Nutzertests mit Rhythmusaufgabe ein positiver, hoher, linearer Zusammenhang.

H4 (Überprüfung der divergenten Validität) Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und den nicht das Ausmaß an intuitiver Benutzung erfassenden divergenten Evaluationsmethoden bestehen bei der Verwendung eines Nutzertests mit Rhythmusaufgabe keine linearen Zusammenhänge.

- **H4.A (Physische Effizienz)** Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und der Anzahl der benötigten Klicks bei der Systemnutzung besteht bei der Verwendung eines Nutzertests mit Rhythmusaufgabe kein linearer Zusammenhang.

- **H4.B (Zeitliche Effizienz)** Zwischen den mit IntuiBeat-S ermittelten Rhythmusabweichungen und der Zeit der motorischen Systemnutzung besteht bei der Verwendung eines Nutzertests mit Rhythmusaufgabe kein linearer Zusammenhang.

5.2.2 Überprüfung der zeitlichen Anwendungseffizienz von IntuiBeat-S

Wie bereits in der Einleitung dieser Arbeit in Kapitel 1 erwähnt, wird im Rahmen des Anwenderprojekts 3D-GUIde eine summative objektive Evaluationsmethode benötigt, die schneller als bekannte objektive Methoden anwendbar ist und somit über eine hohe zeitliche Anwendungseffizienz verfügt, da nur so eine Vielzahl von 3D-CUI-Interaktionslösungen im Projekt in verhältnismäßig kurzer Zeit evaluiert werden können. Da die CHAI-Methode in Abschnitt 3.4 als objektiver Benchmark identifiziert werden konnte, und somit nur diese Methode von bereits bekannten Methoden zur summativen Evaluation von 3D-CUI-Interaktionslösungen im Rahmen des Projekts genutzt werden kann, muss die künftige objektive Benchmarkmethode *IntuiBeat-S* gegenüber der CHAI-Methode schneller anwendbar sein, um eine höhere zeitliche Anwendungseffizienz zu besitzen.

5.2.2.1 Hypothesen

H5 (Überprüfung der zeitlichen Anwendungseffizienz) Bei dem Nutzertest mit Rhythmusaufgabe zeigt sich bei der damit verbundenen künftigen objektiven Benchmarkmethode *IntuiBeat-S* eine höhere zeitliche Anwendungseffizienz als bei dem Nutzertest ohne Rhythmusaufgabe und der damit verbundenen aktuellen objektiven Benchmarkmethode *CHAI*.

5.2.3 Methode

5.2.3.1 Teilnehmer

Für das zweite Experiment wurden 36 Versuchspersonen über das Probandensystem des Instituts für Mensch-Computer-Medien an der Universität Würzburg rekrutiert. Da keine

Datensätze aufgrund fehlender oder unvollständiger Daten ausgeschlossen werden mussten, nahmen 36 Versuchspersonen am Experiment teil, welche alle rechtsfüßig (d.h. der rechte Fuß stellte den dominanten Fuß dar und wurde für die Rhythmuseingabe genutzt) waren. Die Stichprobe setzte sich dabei aus 30 Frauen und sechs Männern zusammen. Das Durchschnittsalter betrug 20.53 Jahre (SD = 1.87). Es handelte sich bei allen Teilnehmern um Studierende der Julius-Maximilians-Universität Würzburg, wovon 12 Personen Mensch-Computer-Systeme (33.30 %) und 24 Personen Medienkommunikation (66.70 %) im Bachelor studierten. Alle Versuchsteilnehmer wurden über das Probanden-System des Instituts Mensch-Computer-Medien rekrutiert und dabei über eine gesonderte Mail darauf hingewiesen, für den Versuch flache Sportschuhe zu tragen, um eine möglichst problemlose Rhythmuseingabe über das USB-Fußpedal zu ermöglichen. Für die Teilnahme an der Untersuchung bekam jede Versuchsperson eine Versuchspersonenstunde gutgeschrieben. Die mit einem TFQ gemessene Vorerfahrung der Versuchspersonen bezüglich der Nutzung von CUIs betrug im Durchschnitt 1.32 (SD = .19) bei einem Maximum von 6 und lag damit, wie bei der Stichprobe erwartet, im unteren Bereich. Alle Versuchspersonen besaßen damit eine geringe Vorerfahrung mit CUIs. Alle Versuchspersonen gaben an, am Experiment freiwillig teilzunehmen.

5.2.3.2 Versuchsdesign

Für die Beantwortung der ersten Forschungsfrage unter Berücksichtigung der Intrusion der Zweitaufgabe wurde ein 2 (Art des Nutzertests) x 2 (unterschiedlich intuitiv gestaltete Software) Mixed Design genutzt. Der erste Faktor *Art des Nutzertests* fungierte dabei als Zwischensubjektfaktor und hatte die Faktorstufen: Nutzertest mit Rhythmusaufgabe vs. Nutzertest ohne Rhythmusaufgabe. Der zweite Faktor *unterschiedlich intuitiv benutzbare Software* fungierte als Innersubjektfaktor und besaß die Faktorstufen: weniger intuitiv benutzbare Software vs. stärker intuitiv benutzbare Software. Die abhängige Variable bildete das *Ausmaß an intuitiver Benutzung.*

Für die Beantwortung des summativen Aspekts der dritten Forschungsfrage wurde ein einfaktorielles experimentelles Between-Subjects Design genutzt. Die unabhängige Variable war die *Art des Nutzertests* mit den Ausprägungen: Nutzertest mit Rhythmusaufgabe vs. Nutzertest ohne Rhythmusaufgabe. Als abhängige Variable fungierte die *zeitliche Anwendungseffizienz* der mit der Art des Nutzertests verbundenen objektiven Benchmarkmethode.

5.2.3.3 Versuchsmaterialien und Maße

Unterschiedlich intuitiv benutzbare Software

Zur Operationalisierung des Innersubjektfaktors *unterschiedlich intuitiv benutzbare Software* wurden auf Basis einer qualitativen Experteneinschätzung ($N_{Experte}$= 5; Vorgehen: siehe entsprechenden Absatz innerhalb des Teilabschnitts 5.1.4.3) das 3D-CUI *Fusion 360* von Autodesk (2017a) als wenig intuitiv benutzbare Software (d.h. erwartete hohe Diversität und Komplexität der Systemnutzung, da das Experten-Tool den Anspruch erhebt, den vollständigen Gestaltungs-, Entwicklungs- und Fertigungsprozess abzubilden) und das

2D-CUI *Affinity Designer* von Serif Inc. (2017) als stärker intuitiv benutzbare Software (d.h. erwartete geringe Diversität und Komplexität, da das Tool sich lediglich auf zweidimensionale Inhalte beschränkt) für die Meta-Evaluation von IntuiBeat-S ausgewählt (siehe Abbildung 5.10). Um diese miteinander vergleichen zu können, wurden verschiedene experimentelle Aufgaben mit der gleichen Zielsetzung wie im ersten Experiment (siehe Abschnitt 5.1) von den gleichen Experten gewählt, die in ein fiktives Szenario eingebettet waren, wo es um das Einfärben von Objekten ging. Die Aufgaben wurden in einer randomisierten Reihenfolge bearbeitet. Da wegen der Anzahl der Freiheitsgrade bei den getesteten 2D-CUI und dem 3D-CUI nicht wie beim Experiment zuvor vollständig identische Aufgaben verwendet werden konnten, kamen stattdessen ähnliche Aufgaben mit vergleichbarer Aufgabenschwierigkeit zum Einsatz. Demzufolge wurde sich zur Erhöhung der Teststärke für einen Vergleich der CUIs als Innersubjektfaktor anstelle eines Zwischensubjektfaktors entschieden, so wie es beim ersten Experiment der Fall war.

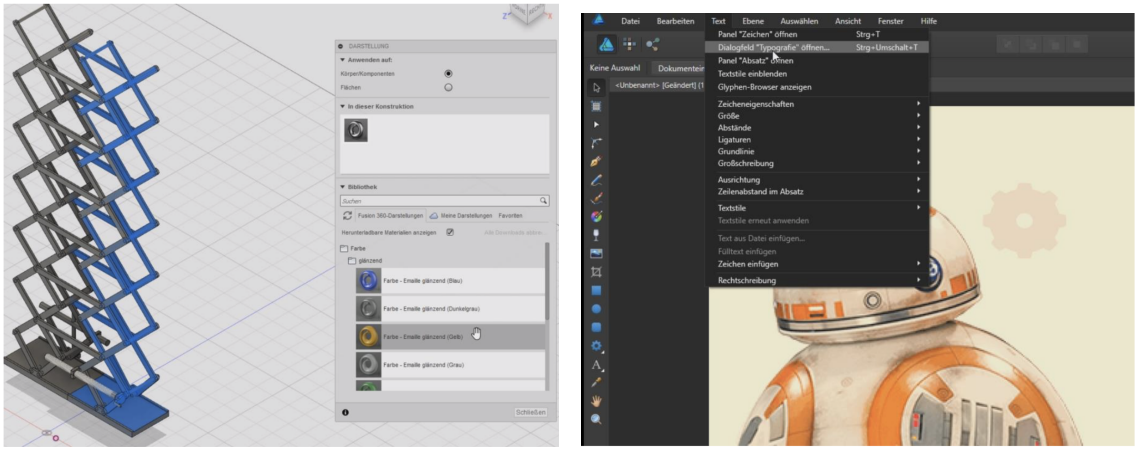


Abbildung 5.8. Die getesteten CUIs *Fusion 360* (links, Autodesk, 2017a) und *Affinity Designer* (rechts, Serif Inc., 2017), welche die beiden Faktorstufen (d.h. Fusion 360: weniger intuitiv benutzbare Software; Affinity Designer: stärker intuitiv benutzbare Software) des Innersubjektfaktors *unterschiedlich intuitiv benutzbare Software* repräsentierten.

Da es sich bei Affinity Designer um ein 2D-CUI handelt, konnte bei beiden CUIs nicht identische experimentelle Testaufgaben und ein identisches Szenario verwendet werden. Es wurde dennoch versucht, bei den gewählten Aufgaben darauf zu achten, dass diese, die von LaViola et al. (2017) beschriebenen domänenübergreifenden, grundlegenden Nutzeraufgaben involvierten, um, wie beim ersten Experiment (siehe Teilabschnitt 5.1.4.3 des ersten Experiments), intuitiv benutzbare CUI-Interaktionslösungen künftig als Interaktionspatterns auf Basis der Ergebnisse ableiten zu können. Des Weiteren wurde bei der konkreten Gestaltung der Färbe-Aufgaben bei beiden CUIs auf eine vergleichbare Schwierigkeit geachtet. Da sich die experimentellen Aufgaben zwar inhaltlich ähnelten, aber nicht identisch waren und damit die Wahrscheinlichkeit für etwaige Übungseffekte als eher gering eingestuft werden konnten, wurden diese als Innersubjektfaktor im Forschungsdesign berücksichtigt. Da jedoch bereits die inhaltliche Ähnlichkeit der CUIs zu Übungseffekten in einem Messwiederholungsdesign führen kann, wurde sich entschieden die Darbietungsreihenfolge der CUIs für jede Versuchsperson zufällig erfolgen zu lassen und die Reihenfolgen über alle Versuchspersonen hinweg auszubalancieren.

Die experimentellen zwei Färbe-Aufgaben (siehe Anhang B.2.1.1), die Versuchspersonen mit Fusion 360 bearbeiten mussten, waren in ein fiktives Anlagenplanungsszenario eines Industriegerüsts eingebettet. Hierbei mussten die Versuchspersonen bei einem Industriegerüst zunächst alle Streben der rechten Seite des Gerüsts gelb einfärben (erste Aufgabe). Im Anschluss sollten die Versuchspersonen die eingefärbten Streben so korrigieren, dass nur die Querstreben des Gerüsts eingefärbt bleiben und die Längsstreben wieder in ihrer grauen Ursprungsfarbe dargestellt werden (zweite Aufgabe). Die experimentellen drei Färbe-Aufgaben (siehe Anhang B.2.1.2), die Versuchspersonen mit Affinity Designer bearbeiten mussten, waren in ein fiktives Gestaltungsszenario einer Geburtstagskarte eingebettet, in dem die Versuchspersonen eine Geburtstagskarte gestalten mussten. Hierzu mussten sie zunächst eine Vorlage laden (erste Aufgabe), diese um drei weitere Objekte (d.h. verschieden große Zahnräder) und den Grußtext „Happy BB-Day“ ergänzen (zweite Aufgabe) und abschließend den Hintergrund der Karte schwarz einfärben (dritte Aufgabe).

Summative Evaluation intuitiver Benutzung mit IntuiBeat-S

Wie bereits im ersten Experiment, stellte das Ausmaß an intuitiver Benutzung auch eine abhängige Variable im zweiten Experiment dar. Als zentrale objektive Methode wurde ebenfalls *IntuiBeat-S* zur Operationalisierung dieser Variable genutzt, da deren wissenschaftliche Güte im Rahmen des zweiten Experiments unter Berücksichtigung der *Art des Nutzertests* untersucht werden sollte, um etwaige Konfundierungen der verwendeten Evaluationsmethoden durch eine mögliche Intrusion der Rhythmuszweitaufgabe ausschließen zu können. Die für die summative Evaluation mit IntuiBeat-S entwickelte Vorgehensweise und die Testumgebung IntuiBeat-Software wurde im zweiten Experiment jedoch nicht verändert und ist somit mit der Beschreibung innerhalb des ersten Experiments identisch (siehe Absatz „Summative Evaluation intuitiver Benutzung mit IntuiBeat-S“ innerhalb des Teilabschnitts 5.1.4.3 des ersten Experiments).

Summative Evaluation intuitiver Benutzung mit anderen Methoden

Wie bereits im ersten Experiment, wurde zur Überprüfung der Intrusion durch die Art des Nutzertests (Hypothese H1), zur Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H2) und zur Überprüfung der konvergenten Validität (Hypothese H3) die abhängige Variable *Ausmaß an intuitiver Benutzung* noch anhand von drei subjektiven Methoden und zwei objektiven Methoden operationalisiert, die als konvergente Quasi-Außenkriterien für die Meta-Evaluation von IntuiBeat-S genutzt wurden. Mit Ausnahme der AV_Effektivität, die im Vergleich zum ersten Experiment nun für eine feinere Erfassung des Anteils erfolgreich durchgeführter Mausklicks an der Gesamtzahl aller getätigten Mausklicks operationalisiert wurde, kamen die gleichen Operationalisierungen und die gleiche Administrationsweise wie im ersten Experiment zum Einsatz, weswegen an dieser Stelle für weitere Informationen auf den entsprechenden Absatz innerhalb des Teilabschnitts 5.1.4.3 des ersten Experiments verwiesen werden soll.

Summative Evaluation von physischer und zeitlicher Effizienz

Wie bereits im ersten Experiment, wurden zur Überprüfung der divergenten Validität (Hypothese H4) noch zwei weitere objektive Methoden als divergente Quasi-Außenkriterien für die Meta-Evaluation von IntuiBeat-S genutzt. Es kamen hierzu die gleichen Operationalisierungen und die gleiche Administrationsweise wie im ersten Experiment zum Einsatz, weswegen an dieser Stelle für weitere Informationen auf den entsprechenden Absatz innerhalb des Teilabschnitts 5.1.4.3 des ersten Experiments verwiesen werden soll.

Bewertung von zeitlicher Anwendungseffizienz

Als Operationalisierung der *zeitlichen Anwendungseffizienz* der, mit der Art des Nutzertests verbundenen, objektiven Benchmarkmethode (d.h. AV_ZeitlicheAnwendungseffizienz) wurde im Rahmen des zweiten Experiments für die Art des summativen Nutzertests (deswegen ohne Berücksichtigung des zusätzlich durchgeführten retrospektiven Interviews, Zeit für das Ausfüllen von Fragebögen und allgemeine Erklärungen bezüglich des Versuchsablaufs) die gesamte Anwendungszeit in Sekunden als Summe aus der Instruktionszeit des Nutzertests (d.h. Zeitabstand zwischen Beginn der Sitzung und Beginn der Aufgabenbearbeitung, welche beim Nutzertest mit der Rhythmusaufgabe die zusätzliche Baseline-Erhebung von IntuiBeat-S und die damit verbundene Instruktion umfasst), der gesamten Durchführungszeit der Systemnutzung (d.h. Zeit für die Aufgabenbearbeitung) und der Auswertungszeit des Nutzertests durch den damit verbundenen objektiven Benchmark (IntuiBeat-S bei Nutzertest mit Rhythmusaufgabe bzw. CHAI bei Nutzertest ohne Rhythmusaufgabe) angegeben.

Da die Auswertung von IntuiBeat-S mithilfe der IntuiBeat-Software durch das Schließen der Software automatisch vorgenommen wird und entsprechende Ergebnisdateien selbsttätig generiert werden (siehe Teilabschnitt 5.1.4.3 des ersten Experiments), wurde die Auswertungszeit für die Methode *IntuiBeat-S* entsprechend mit dem Wert „0“ verrechnet. Die Zeit für die Aufgabenbearbeitung wurde aus den entsprechenden Videoaufzeichnungen (d.h. Länge der Videos) entnommen, da diese ja nur die reine Aufgabenbearbeitung dokumentierten (siehe Teilabschnitt 5.2.3.3 dieses Experiments). Der Beginn der Sitzung wurde aus dem Probandensystem entnommen, da die Versuchspersonen ja einzeln zu verschiedenen Zeitpunkten bestellt wurden. Entsprechende Abweichungen (z.B. Verspätungen) und die Zeiten für allgemeine Einführungen, sowie Fragebogenerhebungen wurden hierbei vom Evaluator notiert und bei der Berechnung berücksichtigt und herausgerechnet. Der Beginn der Baseline-Erhebung zur Kalkulation der Dauer der Instruktion wurde aus der Erstellungszeit der entsprechenden Base-Datei abgeleitet. Es wurde sich bewusst gegen ein Tracking der Zeit durch den Evaluator entschieden, da man die Versuchssituation so real wie möglich gestalten und damit Konfundierungen durch einen etwaigen Präzisionsverlust (z.B. Vergessen rechtzeitig die Zeit zu stoppen) verhindern wollte. Zusätzlich können Folgeexperimente zeitliche Anwendungseffizienz auch ohne großen Aufwand auf diese Weise operationalisieren und ihre Ergebnisse gegebenenfalls sogar post-hoc mit den Ergebnissen dieses Experiments vergleichen, da solche Daten immer verfügbar sind. Eine geringe Anwendungszeit, wie oben beschrieben, indiziert dabei eine hohe zeitliche Anwendungseffizienz des Nutzertests und der damit verbundenen objektiven Benchmarkmethode,

wohingegen eine hohe Anwendungszeit eine geringe Anwendungseffizienz der Art des Nutzertests und der damit verbundenen objektiven Benchmarkmethode indiziert.

Kontrollvariablen und demographische Variablen

Im Rahmen des zweiten Experiments wurden neben den vorgestellten Evaluationsmethoden noch demographische Daten und verschiedene Kontrollvariablen erhoben. Im Gegensatz zum ersten Experiment wurden diese jedoch in Papierform und nicht über LimeSurvey administriert. Die Art der soziodemografischen Daten (d.h. Alter, Geschlecht und Studiengang) unterschied sich hingegen nicht gegenüber dem ersten Experiment (siehe Absatz „Vorerfahrung bei der Nutzung von CUIs“ des Teilabschnitts 5.1.4.3).

Außerdem wurde von den Versuchspersonen die *Vorerfahrung bei der Nutzung von CUIs* als Kontrollvariable mit einem TFQ erhoben (KV_Vorerfahrung), der nun „Affinity Designer“, „Fusion 360“ und „Andere CAD-Software?“ als Items berücksichtigte. Darüber hinaus wurde „Andere CAD-Software?“ als Wildcard genutzt, um die Eingabe eines beliebigen bekannten CUI zu gestatten. Es wurde sich, wie beim ersten Experiment, für die Verwendung des Begriffs „CAD“ anstelle von „CUI“ entschieden, da dieser Begriff in der genutzten Stichprobe wahrscheinlich geläufiger ist und damit keiner Einführung bedarf. Darüber hinaus wurde bei der Konstruktion des TFQ im Vergleich zum ersten Experiment auf die Nennung konkreter, nicht getesteter CUIs als Beispiele verzichtet, da die Expertise der Versuchspersonen mit anderen CUIs mithilfe der Wildcard entsprechend erfasst werden konnte und man bei der Aufführung von Beispielen immer Gefahr läuft, dass diese der Stichprobe nicht geläufig sind. Auf diese Weise lässt sich immer noch die Vorerfahrung mit den getesteten CUIs und bekannten nicht getesteten CUI erfassen, so wie es ein TFQ im Kern vorsieht.

Im Gegensatz zum ersten Experiment (siehe Absatz 5.1.4.3) beurteilten die Versuchspersonen die Items nicht nur bezüglich Nutzungshäufigkeit (Wertebereich: 0 - 6), sondern auch bezüglich des genutzten Funktionsumfangs (Wertebereich: 0 - 4). Im Anschluss wurde dann für jede Dimension (d.h. Nutzungshäufigkeit und Funktionsumfang) ein Mittelwert gebildet und daraufhin ein Gesamtmittelwert über beide Mittelwerte als Operationalisierung der *Vorerfahrung bei der Nutzung von CUIs* berechnet (Wertebereich: 0 - 5). Die Entscheidung für eine Mittelwertbildung anstelle einer Summenberechnung wurde im Rahmen des ersten Experiments ausführlich begründet, weswegen an dieser Stelle lediglich darauf verwiesen wird (siehe Absatz „Vorerfahrung bei der Nutzung von CUIs“ des Teilabschnitts 5.1.4.3). Da kein einheitlicher TFQ existiert und dieser jeweils für die getesteten Softwareanwendungen erstellt werden musste, ist der in diesem Experiment genutzte TFQ vollständig in Anhang B.2.2 dieser Arbeit zu finden.

Apparatur

Die im Rahmen des zweiten Experiments verwendete Apparatur unterschied sich in einigen Punkten von der im ersten Experiment genutzten Apparatur (siehe Absatz „Apparatur“ innerhalb des Teilabschnitts 5.1.4.3 des ersten Experiments). Da der im Rahmen des ersten Experiments genutzte Analyse-PC während des Versuchszeitraums nicht verfügbar war,

wurde stattdessen ein 15 Zoll MacBook Pro (MacOS High Sierra, 2.3 GHz Quad-Core Intel i7, 4 GB Arbeitsspeicher, 1536 MB Intel HD Graphics 4000, Auflösung 2880 x 1800 Pixel) im zweiten Experiment als Analyse-PC genutzt (siehe Abbildung 5.9). Da dieser Analyse-PC zusätzlich nur während des Erhebungszeitraums zur Verfügung stand, wurde für die retrospektive Auswertung der CHAI-Methode der nun wieder verfügbare, im ersten Experiment genutzte Analyse-PC verwendet, auf dem der VLC Mediaplayer in Version 2.2.8 bereits installiert war (siehe Absatz „Apparatur“ innerhalb des Teilabschnitts 5.1.4.3 des ersten Experiments).

Abbildung 5.9. Apparatur des zweiten Experiments, bestehend aus Versuchspersonen-PC (links) und Analyse-PC (rechts).

Darüber hinaus kam im zweiten Experiment kein dedizierter Evaluatoren-PC zum Einsatz, da als Bildschirmaufzeichnungssoftware anstelle von Morae die quelloffene Bildschirmaufzeichnungssoftware Captura von Sachin (2017) verwendet wurde, bei der die Aufzeichnung nur lokal über den Versuchspersonen-PC gestartet wird und diese damit keine Fernsteuerung von einem anderen PC zulässt. Es wurde sich für eine andere Software entschieden, da die von Morae angebotene Fernsteuerung zwar einen erhöhten Komfort bei der Testdurchführung bietet, sich dieser aber in den für das Experiment genutzten kleinen Örtlichkeiten nicht wirklich auswirkt und den damit verbundenen komplexen Aufbau mit drei verschiedenen Rechnern rechtfertigt. Captura bietet genauso wie Morae die Möglichkeit Nutzereingaben wie beispielsweise Mausklicks mitzuprotokollieren, was für die Anwendung einiger Evaluationsmethoden (z.B. CHAI-Methode) benötigt wurde. Captura wurde vom Evaluator im Vorfeld der Untersuchung auf dem Versuchspersonen-PC installiert.

Die räumliche Anordnung der Apparatur musste im Vergleich zum ersten Experiment geändert werden, da das im ersten Experiment verwendete Labor im Untersuchungszeitraum nicht zur Verfügung stand. Der Evaluator nahm deswegen nun rechts von dem Analyse-PC Platz (siehe Abbildung 5.9). Als Versuchspersonen-PC fungierte das gleiche MSI GT62VR Gaming Notebook wie im ersten Experiment (siehe Absatz „Apparatur“ innerhalb des Teilabschnitts 5.1.4.3). Die eingebaute Tastatur und das Trackpad des Notebooks fungierten hier ebenfalls nicht als Eingabegeräte. Stattdessen wurde auf eine externe kabelgebundene Maus M-UAV-DEL8 von Dell und eine externe kabelgebundene Tastatur KB520 von Fujitsu zurückgegriffen. Als Bildschirm kam ein dedizierter 24 Zoll LCD-Bildschirm

B24W-6 von Fujitsu (Auflösung 1920 x 1200 Pixel) zum Einsatz. Der Abstand zwischen den Versuchspersonen und dem Bildschirm betrug hier, wie im ersten Experiment, circa 70 cm. Der eingebaute 15 Zoll Bildschirm des Gaming Notebooks wurde während der Erhebung zugeklappt, um Ablenkungen zu vermeiden (siehe Abbildung 5.9). Im Gegensatz zum ersten Experiment kam anstelle eines Drehstuhls ein gewöhnlicher Stuhl als Sitzgelegenheit für die Versuchsperson zum Einsatz, da im ersten Experiment bei der Ausrichtung des Fußpedals die Höhe des Stuhls nur selten verstellt werden musste. Die als Faktorstufen des Innersubjektfaktors fungierenden CUIs wurden, wie im ersten Experiment, auf dem Versuchspersonen-PC installiert. Der Austausch der Aufzeichnungsdateien zwischen dem Versuchspersonen-PC und dem Analyse-PC erfolgte über ein gemeinsames Netzlaufwerk.

5.2.3.4 Versuchsdurchführung

Wie im ersten Experiment erfolgte auch beim zweiten Experiment die Rekrutierung der Versuchspersonen über das Probanden-System des Instituts für Mensch-Computer-Medien der Universität Würzburg. Das Experiment wurde darin als „Nutzertest verschiedener Software-Anwendungen“ beworben und identisch wie im ersten Experiment beschrieben (siehe Teilabschnitt 5.1.4.4). Alle angebotenen Termine waren ebenfalls Einzeltermine, da auch immer nur eine Person gleichzeitig getestet wurde.

Das zweite Experiment wurde in einem kleinen, reizabgeschirmten, stimulusarmen Labor des Lehrstuhls für Psychologische Ergonomie der Universität Würzburg durchgeführt, welches aber etwas kleiner als das im ersten Experiment verwendete Labor war (siehe Abbildung 5.9). Der Geräuschpegel von außen war über den gesamten Erhebungszeitraum konstant niedrig. Bevor die Versuchspersonen das Testlabor betreten durften, wurden sie vom Evaluator mündlich aufgefordert, ihre Mobiltelefone in den Flugmodus zu schalten, um eine Ablenkung durch diese während der Untersuchung zu vermeiden. Die eigentliche Datenerhebung der Studie unterteilte sich in drei Phasen (d.h. Vorphase, Hauptphase, Abschlussphase), die insgesamt circa 50 Minuten in Anspruch nahmen. Im Folgenden soll nur auf die Unterschiede in den einzelnen Phasen der Versuchsdurchführung im Vergleich zum ersten Experiment eingegangen werden und für weitere Details bezüglich der Gemeinsamkeit auf Teilabschnitt 5.1.4.4 des ersten Experiments verwiesen werden.

Vorphase (5 Minuten)

Im Vergleich zum ersten Experiment, kam in der Vorphase des zweiten Experiments ein anderer TFQ zum Einsatz. Es wurde außerdem kein PROMS zur Erfassung der rhythmischen Wahrnehmung administriert und den Versuchspersonen alle Materialien der Vorphase in Papierform vorgelegt. Die restliche Vorphase wurde im Vergleich zum ersten Experiment identisch gestaltet, weswegen für genauere Details auf den Teilabschnitt 5.1.4.4 des ersten Experiments verwiesen wird.

Hauptphase (40 Minuten)

Da im Gegensatz zum ersten Experiment die Art des Nutzertests bei der Überprüfung der wissenschaftlichen Güte von IntuiBeat-S berücksichtigt werden sollte, wurden die Versuchspersonen zu Beginn der Hauptphase den Versuchsbedingungen *Nutzertest mit Rhythmusaufgabe* und *Nutzertest ohne Rhythmusaufgabe* randomisiert zugewiesen und diese Zuweisung bezüglich der *Art des Nutzertests* (d.h. Zwischensubjektfaktor) über alle Versuchspersonen hinweg ausbalanciert. Versuchspersonen in der Versuchsbedingung *Nutzertest ohne Rhythmusaufgabe* erhielten zunächst eine standardisierte mündliche Instruktion durch den Evaluator für die Hauptphase der Untersuchung. Im Rahmen dieser mündlichen Instruktion wurde den Versuchspersonen zunächst der genaue Ablauf des Nutzertests mitgeteilt und ihnen dabei erläutert, dass das Ziel eines solchen Tests darin bestehe, sowohl Probleme des Systems zu identifizieren, als auch das Ausmaß an intuitiver Benutzung der CUIs zu bewerten. Daraufhin wurde den Versuchspersonen randomisiert das erste der beiden CUIs vom Evaluator standardisiert mündlich vorgestellt und ihnen das dazugehörige Szenario präsentiert. Auf diese Weise wurden die Versuchspersonen auch den Ausprägungen des Innersubjektfaktors *weniger intuitiv benutzbare Software* oder *stärker intuitive benutzbare Software* zugewiesen und diese Zuweisung über alle Versuchspersonen hinweg ausbalanciert.

Vor Beginn der eigentlichen Aufgabenbearbeitung händigte der Evaluator den Versuchspersonen die Instruktion der ersten Testaufgabe des ersten CUIs und das damit verbundene Szenario in Papierform aus. Daraufhin öffnete der Evaluator die entsprechende Software und wählte beim 3D-CUI *Fusion 360* zusätzlich die 3D-Szene des dazugehörigen CUI-Szenarios im Vorfeld aus (Datei „FusionExperiment2.f3d“), da es beim 2D-CUI *Affinity Designer* keine 3D-Szene gab und die Probanden stattdessen ein Bild (Datei „BB8.png“) als erste Aufgabe in das Programm laden mussten. Nachdem die Versuchspersonen dem Evaluator durch ein Handzeichen signalisiert hatten, dass sie mit der Bearbeitung der ersten Aufgabe beginnen möchten, wies der Evaluator auf die stillschweigende Bearbeitung der Aufgabe hin und startete daraufhin die Bildschirmaufzeichnung (inkl. Aufzeichnung von Mausbewegungen). Nach der Erledigung der ersten Aufgabe des ersten CUI gaben die Versuchspersonen dem Evaluator erneut ein Handzeichen und dieser stoppte daraufhin die Aufzeichnung. Da alle Testaufgaben so ausgewählt wurden, dass ein Expertennutzer eine Testaufgabe in einer Minute erledigen kann, wurde die Aufgabenbearbeitung vom Evaluator nach fünf Minuten abgebrochen. Nachdem der Evaluator die Aufzeichnungen gestoppt hatte, instruierte er die Versuchsteilnehmer anschließend die SEA-Skala und den NASA-RTLX auszufüllen, welche ihnen zu diesem Zweck in Papierform ausgehändigt wurden. Die Darbietungsreihenfolge der SEA-Skala und des NASA-RTLX wurden über alle Versuchspersonen hinweg vollständig ausbalanciert. Nachdem die Versuchspersonen die beiden Fragebögen ausgefüllt hatten, händigte der Evaluator ihnen die Instruktion der zweiten Aufgabe in Papierform aus und bat sie die Instruktion aufmerksam zu lesen. Die geschilderte Bearbeitungsprozedur wurde für diese und die restlichen Testaufgaben wiederholt.

Als die Versuchspersonen alle Testaufgaben des ersten CUI absolviert hatten, wurden sie vom Evaluator standardisiert mündlich aufgefordert die von ihnen getestete Softwareanwendung mit dem QUESI zu bewerten, der ihnen ebenfalls in Papierform ausgehändigt

wurde. Im Anschluss folgte ein retrospektives Interview, um Nutzungsprobleme zu identifizieren. Nach dem ersten Interview wurde den Versuchspersonen vom Evaluator das zweite CUI standardisiert mündlich vorgestellt und das entsprechende Szenario in Papierform präsentiert. Die Vorgehensweise bei der Bearbeitung des zweiten CUI und des anschließenden retrospektiven Interviews gestaltete sich identisch wie beim ersten CUI. Mit Abschluss des letzten Interviews endete die Hauptphase der Versuchsbedingung *Nutzertest ohne Rhythmusaufgabe*.

Versuchspersonen in der Versuchsbedingung *Nutzertest mit Rhythmusaufgabe* erhielten eine ähnliche standardisierte mündliche Instruktion durch den Evaluator wie Versuchspersonen in der Versuchsbedingung *Nutzertest ohne Rhythmusaufgabe*. Wie beim ersten Experiment (siehe Teilabschnitt 5.1.4.4), wurde ihnen vom Evaluator noch zusätzlich mitgeteilt, dass die intuitive Benutzung neben anderen Methoden (z.B. Fragebögen) auch mithilfe einer Rhythmuszweitaufgabe gemessen werden soll. Dementsprechend wurde von Versuchspersonen daraufhin eine Rhythmus-Baseline erhoben, wie es im Rahmen des ersten Experiments beschrieben wurde. Für genauere Details wird auf den Absatz „Funktionsweise des Baseline-Modus der IntuiBeat-Software und Beschreibung des Programmablaufs“ des vorigen Teilabschnitts 5.1.4.3 verwiesen. Die Vorgehensweise bei der Bearbeitung der CUIs und der retrospektiven Interviews gestaltete sich prinzipiell identisch wie in der Versuchsbedingung *Nutzertest ohne Rhythmusaufgabe*. Der einzige Unterschied bei der Bearbeitung der Aufgaben mit den CUIs bestand in dieser Bedingung darin, dass Versuchspersonen instruiert wurden, parallel zur Systemnutzung den gelernten Rhythmus mit einem USB-Fußpedal zu klopfen. Der Evaluator musste deswegen neben der Aufzeichnung des Bildschirms und der Mauswege (über den Versuchspersonen-PC) auch gleichzeitig die Aufzeichnung des geklopften Rhythmus zu Beginn einer jeden Testaufgabe starten und zum Ende jeder Testaufgabe beenden. Im Zuge dessen sicherte der Evaluator die erhobenen Daten entsprechend. Für genauere Details wird auf den Absatz „Funktionsweise des Experimental-Modus der IntuiBeat-Software und Beschreibung des Programmablaufs“ des vorigen Teilabschnitts 5.1.4.3 verwiesen. Nachdem die Versuchspersonen auch in dieser Versuchsbedingung die Bearbeitung von beiden CUIs und die damit verbundenen retrospektiven Interviews abgeschlossen hatten, endete die Hauptphase der Versuchsbedingung *Nutzertest mit Rhythmusaufgabe*. Die Hauptphase des Experiments war in dieser Versuchsbedingung fünf Minuten länger, weil diese Zeit für das Erheben der individuellen Baseline veranschlagt wurde.

Abschlussphase (5 Minuten)

Bei der Abschlussphase bestand sowohl in der Dauer als auch in der Durchführung kein Unterschied zwischen dem ersten und zweiten Experiment (siehe Teilabschnitt 5.1.4.4).

5.2.3.5 Statistische Auswertung

Die statistische Datenauswertung erfolgte mittels IBM SPSS Statistics 25 für macOS. Zunächst wurde vor der eigentlichen Datenauswertung geprüft, ob die Daten Ausreißer enthielten, die Datensätze der erhobenen Stichprobe vollständig und die Voraussetzungen

der statistischen Tests erfüllt waren. Alle statistischen Tests und Analysen der Teststärke erfolgten zweiseitig.

Überprüfung der Validität von IntuiBeat-S

Zunächst erfolgte die Bestimmung der Konstruktvalidität von IntuiBeat-S unter Berücksichtigung der Art der als Basis für die Datenauswertung verwendeten Nutzertests (siehe Teilabschnitt 5.2.4.2). Das mehrstufige Vorgehen unterschied sich dabei kaum vom Vorgehen des ersten Experiments (siehe Teilabschnitt 5.1.4.5). Die statistische Auswertung erfolgte lediglich mit einem zusätzlichen Schritt zur Überprüfung der Intrusion durch die Art des Nutzertests (Hypothese H1). Die anderen beiden Schritte Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H2) und die Überprüfung der konvergenten und divergenten Validität von IntuiBeat-S (Hypothese H3 und H4) blieben gegenüber dem ersten Experiment identisch (siehe Teilabschnitt 5.1.4.5).

Als erster Schritt wurde zunächst die **mögliche Intrusion der Art des Nutzertests** (Hypothese H1) multivariat in einem zweifaktoriellen Mixed Design überprüft (siehe Teilabschnitt 5.2.4.2). Der Zwischensubjektfaktor war hier die unabhängige Variable *Art des Nutzertests* mit den Faktorstufen: Nutzertest mit Rhythmusaufgabe vs. Nutzertest ohne Rhythmusaufgabe. Als Innersubjektfaktor fungierte die unabhängige Variable *unterschiedlich intuitiv benutzbare Software* mit den Faktorstufen: weniger intuitiv benutzbare Software vs. stärker intuitiv benutzbare Software. Als abhängige Variablen kamen die als konvergente Quasi-Außenkriterien genutzten Evaluationsmethoden, die das *Ausmaß an intuitiver Benutzung* auf unterschiedliche Art und Weise operationalisieren, zum Einsatz. Da die *Vorerfahrung bei der Nutzung von CUIs*, die als Kontrollvariable KV_Vorerfahrung erhoben wurde, einen ungewollten Einfluss auf den Vergleich der beiden Faktorstufen des Zwischensubjektfaktors haben könnte, musste diese mögliche Konfundierung durch die *Vorerfahrung bei der Nutzung von CUIs* zuerst ausgeschlossen werden. Hierzu wurde ein t-Test für unabhängige Stichproben bezüglich der Vorerfahrung bei der Nutzung von CUIs, sowie Pearson-Produkt-Moment-Korrelationen zwischen KV_Vorerfahrung und den, das Ausmaß an intuitiver Benutzung indizierenden, konvergenten Evaluationsmethoden innerhalb beider Faktorstufen des Zwischensubjektfaktors berechnet. Das Signifikanzniveau betrug hier ebenfalls $\alpha = .05$. Die Effektstärke für den t-Test wurde mithilfe von G*Power (Faul et al., 2009) berechnet und als Absolutwert interpretiert. Dabei indizierten die Effektstärken d um .20 einen kleinen, um .50 einen mittleren und um .80 einen großen Effekt (J. Cohen, 1992). Ein Korrelationskoeffizient $|r|$ um .10 wurde als ein schwacher, ein Korrelationskoeffizient $|r|$ um .30 als ein moderater, und ein Korrelationskoeffizient $|r|$ um .50 als hoher Zusammenhang interpretiert (J. Cohen, 1988). Die Überprüfung der statistischen Voraussetzungen der Pearson-Produkt-Moment-Korrelationen und des t-Tests werden im folgenden Ergebnisteil berichtet (siehe Teilabschnitt 5.2.4.1).

Da im Zuge dieser Analyse keine signifikanten Unterschiede und Zusammenhänge festgestellt werden konnten (d.h. kein ungewollter Einfluss der Kontrollvariable auf die abhängigen Variablen), wurde von einer multivariaten Kovarianzanalyse (MANCOVA) mit Messwiederholung abgesehen (Döring & Bortz, 2016; Field, 2017). Es wurde sich stattdessen für eine Datenauswertung mithilfe einer multivariaten Varianzanalyse (MANOVA) mit Messwiederholung ohne Berücksichtigung der Kontrollvariable zur Überprüfung der

Intrusion durch die Art des Nutzertests (Hypothese H1) entschieden (siehe Teilabschnitt 5.2.4.2). Das Signifikanzniveau für alle damit verbundenen statistischen Berechnungen betrug $\alpha = .05$ und die mithilfe von IBM SPSS Statistics 25 berechneten Effektgrößen wurden als Absolutwerte berichtet und interpretiert. Bei der multivariaten Varianzanalyse wurde eine absolute Effektstärke η_p^2 um .01 als klein, eine Effektstärke η_p^2 um .06 als mittel und eine Effektstärke η_p^2 um .14 als groß interpretiert (J. Cohen, 1988). Die Überprüfung der statistischen Voraussetzungen der MANOVA wird im folgenden Ergebnisteil berichtet (siehe Teilabschnitt 5.2.4.1).

Da mit dieser multivariaten Auswertung erwartungsgemäß ein signifikanter Haupteffekt des Innersubjektfaktors (H1.A), jedoch kein signifikanter Haupteffekt des Zwischensubjektfaktors (H1.B), sowie kein signifikanter Interaktionseffekt (H1.C) beobachtet werden konnte, erfolgte in einem zweiten Schritt die **Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs** (Hypothese H2) mithilfe eines einfaktoriellen Within-Subjects Designs innerhalb der Bedingung *Nutzertest mit Rhythmusaufgabe* des Zwischensubjektfaktors (siehe Teilabschnitt 5.2.4.2). Die unabhängige Variable war die *unterschiedlich intuitiv benutzbare Software* mit den Faktorstufen: weniger intuitiv benutzbare Software vs. stärker intuitiv benutzbare Software. Als abhängige Variablen fungierten neben IntuiBeat-S auch die anderen konvergenten summativen Evaluationsmethoden, die das *Ausmaß an intuitiver Benutzung* auf unterschiedliche Art und Weise operationalisierten. Bevor jedoch die einzelnen Hypothesen durch mehrere univariate Varianzanalysen (ANOVAs) mit Messwiederholung überprüft wurden, wurde eine multivariate Varianzanalyse (MANOVA) mit Messwiederholung innerhalb der Bedingung *Nutzertest mit Rhythmusaufgabe* des Zwischensubjektfaktors in Übereinstimmung mit Cramer und Bock (1966) berechnet, da diese dabei helfen kann, den Alphafehler bei darauffolgenden ANOVAs mit Messwiederholung zu kontrollieren, ohne dass man das Signifikanzniveau korrigieren muss, wie es bei mehreren ANOVAS normalerweise nötig wäre (Field, 2017).

Aufgrund der geringen Stichprobengröße im Verhältnis zur hohen Anzahl an konvergenten Evaluationsmethoden würde eine Bonferroni-Korrektur oder die Korrektur mithilfe eines ähnlichen Verfahrens (z.B. Bonferroni-Holm, Hochberg) sehr zu Lasten der Teststärke gehen (Field, 2017). Das Signifikanzniveau für alle damit verbundenen statistischen Berechnungen betrug $\alpha = .05$ und die mithilfe von IBM SPSS Statistics 25 berechneten Effektgrößen wurden als Absolutwerte berichtet und interpretiert. Bei der multivariaten Varianzanalyse wurde eine absolute Effektstärke η_p^2 um .01 als klein, eine Effektstärke η_p^2 um .06 als mittel und eine Effektstärke η_p^2 um .14 als groß interpretiert (J. Cohen, 1988). Einer signifikanten MANOVA folgten ANOVAS zur Überprüfung, ob die einzelnen Evaluationsmethoden wirklich in der Lage sind, zwischen den Faktorstufen des Innersubjektfaktors bezüglich des Ausmaßes an intuitiver Benutzung differenzieren zu können (H2.A bis H2.F). Bei allen univariaten Varianzanalysen wurde eine absolute Effektstärke η^2 um .01 als klein, eine Effektstärke η^2 um .06 als mittel und eine Effektstärke η^2 um .14 als groß interpretiert (J. Cohen, 1988). Die Überprüfung der statistischen Voraussetzungen der MANOVA und den darauffolgenden ANOVAs werden im folgenden Ergebnisteil berichtet (siehe Teilabschnitt 5.2.4.1).

Nachdem, wie bereits im ersten Experiment (siehe Teilabschnitt 5.1), der erste Teil des Vorgehens zum Nachweis der Konstruktvalidität von IntuiBeat-S abgeschlossen und damit

abgesichert war, dass der Zwischensubjektfaktor, die Art des als Basis für die konvergenten Evaluationsmethoden genutzten Nutzertests, keine Auswirkung auf die Beurteilung des Ausmaßes an intuitiver Benutzung durch die konvergenten Evaluationsmethoden hatte, und alle konvergenten Evaluationsmethoden die Unterschiede in der intuitiver Gestaltung der beiden CUIs feststellen konnten, erfolgte wie im ersten Experiment noch die Überprüfung der konvergenten und divergenten Validität von IntuiBeat-S als summative Evaluationsmethode für intuitive Benutzung bei der Anwendung eines Nutzertests mit Rhythmusaufgabe (Hypothesen H3 und H4). Als erster Schritt wurde hier zunächst die **konvergente Validität von IntuiBeat-S** (Hypothese H3) bestimmt (siehe Teilabschnitt 5.2.4.2), indem Pearson-Produkt-Moment-Korrelationen zwischen allen konvergenten Evaluationsmethoden für beide Stufen des Innersubjektfaktors unter der Bedingung *Nutzertest mit Rhythmusaufgabe* des Zwischensubjektfaktors berechnet wurden. Das Signifikanzniveau betrug hier $\alpha = .05$. Dabei wurde geprüft, ob, wie erwartet, hohe signifikante Zusammenhänge zwischen IntuiBeat-S und den konvergenten Evaluationsmethoden bestehen (H3.A bis H3.E). Als nächster Schritt wurde die **divergente Validität von IntuiBeat-S** (Hypothese H4) bestimmt (siehe Teilabschnitt 5.2.4.2), indem ebenfalls Pearson-Produkt-Moment-Korrelationen zwischen IntuiBeat-S und allen divergenten Evaluationsmethoden, die nicht das Ausmaß an intuitiver Benutzung abbilden, innerhalb beider Stufen des Innersubjektfaktors unter der Bedingung *Nutzertest mit Rhythmusaufgabe* des Zwischensubjektfaktors berechnet wurden. Das Signifikanzniveau betrug hier ebenfalls $\alpha = .05$. Dabei wurde geprüft, ob, wie erwartet, keine signifikanten Zusammenhänge zwischen IntuiBeat-S und den anderen nicht intuitive Benutzung evaluierenden divergenten Methoden bestehen (H4.A und H4.B). Zur Überprüfung der konvergenten und divergenten Validität wurden in Anlehnung an J. Cohen (1988) ein Korrelationskoeffizient $|r|$ um .10 als ein schwacher, ein Korrelationskoeffizient $|r|$ um .30 als ein moderater, und ein Korrelationskoeffizient $|r|$ um .50 als hoher Zusammenhang interpretiert. Die Überprüfung der statistischen Voraussetzungen der Pearson-Produkt-Moment-Korrelationen werden im folgenden Ergebnisteil berichtet (siehe Teilabschnitt 5.2.4.1).

Überprüfung der zeitlichen Anwendungseffizienz von IntuiBeat-S

Die Überprüfung der zeitlichen Anwendungseffizienz von IntuiBeat-S (Hypothese H5) und damit die Beantwortung des summativen Aspekts der dritten Forschungsfrage erfolgte im einfaktoriellen Between-Subjects Design (siehe Teilabschnitt 5.2.4.3). Als unabhängige Variable fungierte die *Art des Nutzertests* mit den Ausprägungen: Nutzertest mit Rhythmusaufgabe vs. Nutzertest ohne Rhythmusaufgabe. Als abhängige Variable diente die *zeitliche Anwendungseffizienz* der mit der Art des Nutzertests verbundenen objektiven Benchmarkmethode. Das Signifikanzniveau für alle damit verbundenen statistischen Berechnungen betrug $\alpha = .05$ und die berechneten Effektgrößen wurden als Absolutwerte berichtet. Die Datenauswertung erfolgte mithilfe eines t-Tests für unabhängige Stichproben. Dabei indizierten absolute Effektstärken d um .20 einen kleinen, um .50 einen mittleren und um .80 einen großen Effekt (J. Cohen, 1992). Die Effektstärke für den t-Test wurde mithilfe von G*Power (Faul et al., 2009) berechnet und als Absolutwerte interpretiert. Die Überprüfung der statistischen Voraussetzungen des t-Tests werden im folgenden Ergebnisteil berichtet (siehe Teilabschnitt 5.2.4.1).

5.2.4 Ergebnisse

Im folgenden Abschnitt werden die Ergebnisse bezüglich der in den Teilabschnitten 5.2.1.1 und 5.2.2 beschriebenen Hypothesen deskriptiv und inferenzstatistisch berichtet. Vor der eigentlichen Datenanalyse wird zunächst auf die Überprüfung der statistischen Voraussetzungen eingegangen. Dabei wurde bei allen abhängigen Variablen und Kontrollvariablen ein metrisches Skalenniveau angenommen.

5.2.4.1 Überprüfung der statistischen Voraussetzungen

Überprüfung von Ausreißern

Zur Überprüfung univariater Ausreißer wurden vor der Untersuchung der einzelnen Hypothesen modifizierte z-Werte herangezogen, da z-Werte generell ein häufig angewendetes Verfahren zur Identifikation univariater Ausreißer darstellen (Cousineau & Chartier, 2010; Shiffler, 1988). Es wurde sich speziell für die Verwendung von modifizierten z-Werten entschieden, da diese auch bei Stichproben mit geringer Größe mit hoher Zuverlässigkeit funktionieren (Garcia, 2012; Iglewicz & Hoaglin, 1993; Seo, 2006). Als univariate Ausreißer wurden in Anlehnung an Iglewicz und Hoaglin (1993) Werte identifiziert, deren absoluter modifizierter z-Wert größer als 3.5 lag. Es hätte auf diese Weise drei Werte bei der AV_IntuiBeat-S (weniger intuitiv benutzbare Software: $z = 5.09$, $z = 5.82$; stärker intuitiv benutzbare Software: $z = 4.03$) ausgeschlossen werden müssen, worauf aufgrund der kleinen Stichprobe ($N = 36$; $n = 18$ bei der Verwendung eines Nutzertests mit Rhythmusaufgabe) verzichtet wurde.

Zur Überprüfung multivariater Ausreißer wurde vor Berechnung der multivariaten Varianzanalysen zur Überprüfung der Intrusion durch die Art des Nutzertests (Hypothese H1) und zur Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H2) die Mahalanobis-Distanz herangezogen, da diese ein häufig angewendetes Verfahren zur Identifikation multivariater Ausreißer darstellt (De Maesschalck et al., 2000; Hadi, 1992). Anhand eines Vergleichs der Mahalanobis-Distanz mit den kritischen Werten der Chi-Quadrat-Verteilung (mit $\alpha = .001$ und bei der Gleichsetzung der Freiheitsgrade mit der Anzahl der konvergenten Quasi-Außenkriterien zur Überprüfung der Hypothese H1 bzw. mit der Anzahl aller Evaluationsmethoden inkl. IntuiBeat-S, die das Ausmaß an intuitiver Benutzung in diesem Experiment bei der Verwendung eines Nutzertests mit Rhythmusaufgabe zur Überprüfung der Hypothese H2 operationalisierten), wurde festgestellt, dass sich keine multivariaten Ausreißer in den Daten befanden.

Überprüfung der Voraussetzung der Normalverteilung

Die univariate Normalverteilung der abhängigen Variablen und Kontrollvariablen wurde für die, zur Beantwortung der jeweiligen Forschungsfragen benötigten Kombinationen, des Zwischensubjektfaktors und des Innersubjektfaktors (siehe Teilabschnitt 5.2.3.5) mittels Kolmogorov-Smirnov-Tests ($p \geq .05$, siehe Field, 2017) und Sichtprüfung anhand eines Q-Q-Diagramms geprüft.

Bezüglich der ersten Forschungsfrage konnten dabei auf Basis dieser beiden Kriterien bei den Variablen AV_IntuiBeat-S, AV_CHAI, KV_Vorerfahrung, AV_PhysischeEffizienz und AV_Effektivität keine Normalverteilung im Rahmen der Überprüfung der Intrusion durch die Art des Nutzertests (Hypothese H1) festgestellt werden. Innerhalb der Faktorstufe *Nutzertest mit Rhythmusaufgabe* des Zwischensubjektfaktors konnte bei diesen Variablen ebenfalls keine Normalverteilung im Rahmen der Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H2) und der Überprüfung der konvergenten und divergenten Validität (Hypothesen H3 und H4) festgestellt werden. Da eine univariate Varianzanalyse mit Messwiederholung bei etwa gleich großen Gruppen robust gegenüber Verletzungen der Normalverteilungsannahme ist (Blanca et al., 2017; Glass et al., 1972; Lix et al., 1996; Schmider et al., 2010), insbesondere wenn keine weiteren Annahmen verletzt wurden (Berkovits, Hancock, & Nevitt, 2000), so wie es beim zweiten Experiment der Fall war (siehe Überprüfung der weiteren Voraussetzungen), wurde bei der univariaten varianzanalytischen Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H2) nicht auf nonparametrische Verfahren zurückgegriffen. Aufgrund der Tatsache, dass auch die zur Überprüfung der konvergenten und divergenten Validität (Hypothesen H3 und H4) eingesetzten Pearson-Produkt-Moment-Korrelationen (Edgell & Noon, 1984; Havlicek & Peterson, 1976; Liddell & Kruschke, 2018), sowie der hier zur Überprüfung der Signifikanz genutzte gepaarte t-Test im Allgemeinen robust gegenüber Verletzungen der Normalverteilungsannahme sind (Glass et al., 1972; Pagano, 2006; Salkind, 2010; Wilcox, 2011), wurde sich an dieser Stelle, für eine eindeutige Interpretation (d.h. Mittelwerte statt Mediane) und Vergleichbarkeit der Ergebnisse der verschiedenen Experimente, gegen Transformationen und entsprechende nonparametrische Verfahren zur Überprüfung der konvergenten bzw. divergenten Validität (Hypothesen H3 und H4) entschieden.

Bezüglich des summativen Aspekts der dritten Forschungsfrage konnte auf Basis dieser beiden Kriterien der AV_ZeitlicheAnwendungseffizienz keine Normalverteilung im Rahmen der Überprüfung der zeitlichen Anwendungseffizienz (Hypothese H5) festgestellt werden. Aufgrund der Tatsache, dass der zur Überprüfung der Signifikanz genutzte, ungepaarte t-Test im Allgemeinen robust gegenüber Verletzungen der Normalverteilungsannahme ist (Glass et al., 1972; Pagano, 2006; Salkind, 2010; Wilcox, 2011), wurde sich wie bereits im Rahmen der Beantwortung der ersten Forschungsfrage, für eine eindeutige Interpretation (d.h. Mittelwerte statt Mediane) und Vergleichbarkeit der Ergebnisse gegenüber Folgeexperimenten, gegen Transformationen und entsprechende nonparametrische Verfahren zur Überprüfung der zeitlichen Anwendungseffizienz (Hypothese H5) entschieden.

Aufgrund der Tatsache, dass mit dem für die statistische Auswertung genutzten, IBM SPSS Statistics 25 kein multivariater Shapiro-Wilk-Test berechnet werden kann, wurde eine Überprüfung der multivariaten Normalverteilung der abhängigen Variablen und Kontrollvariablen lediglich näherungsweise durch Überprüfung der univariaten Normalverteilung vorgenommen, so wie es eben beschrieben wurde. Da die Normalverteilungsannahme bei einigen Variablen verletzt war, erfolgten die multivariate Varianzanalysen unter Verwendung der Pillai-Spur zur Überprüfung der Intrusion durch die Art des Nutzertests (Hypothese H1) und zur Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothesen H2), da diese nicht nur als robust gegenüber Modellverletzungen gilt, sondern auch eine große statistische Power bei kleinen Stichproben besitzt (Tabachnick et al., 2007).

Überprüfung der Voraussetzung von Linearität und Sicherstellung fehlender Multikollinearität

Die Linearität der abhängigen Variablen wurde im Rahmen der Überprüfung der Intrusion durch die Art des Nutzertests (Hypothese H1) und zur Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H2), sowie der Überprüfung der konvergenten und divergenten Validität (Hypothesen H3 und H4) für jede Kombination des Zwischensubjektfaktors und des Innersubjektfaktors (Hypothese H1) bzw. innerhalb der Faktorstufe *Nutzertest mit Rhythmusaufgabe* für jede Faktorstufe des Innersubjektfaktors (Hypothesen H2, H3, H4) univariat anhand einer Streudiagrammmatrix überprüft. Dabei wurde keine Verletzung dieser Voraussetzung festgestellt. Die, für die multivariaten Varianzanalysen zur Überprüfung der Intrusion durch die Art des Nutzertests (Hypothese H1) und zur Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H2) außerdem erforderliche, fehlende Multikollinearität wurde jeweils mithilfe von Pearson-Produkt-Moment-Korrelationen zwischen den abhängigen Variablen für jede Kombination des Zwischensubjektfaktors und des Innersubjektfaktors (Hypothese H1) bzw. innerhalb der Faktorstufe *Nutzertest mit Rhythmusaufgabe* für jede Faktorstufe des Innersubjektfaktors (Hypothese H2) überprüft. Da alle Korrelationen nicht übermäßig hoch lagen ($|r| < .8$), wurde Multikollinearität ausgeschlossen (Berry et al., 1985; Rockwell, 1975). Es wurden an dieser Stelle trotzdem Pearson-Produkt-Moment-Korrelationen berechnet, obwohl die Normalverteilungsannahme einiger Variablen (siehe „Überprüfung der Voraussetzung der Normalverteilung“ des Teilabschnitts 5.2.4.1) verletzt war, da diese bei etwa gleich großen Gruppen gegenüber dieser Verletzung relativ robust reagieren (Edgell & Noon, 1984; Havlicek & Peterson, 1976; Liddell & Kruschke, 2018).

Überprüfung der Homoskedastizität, der Homogenität der Varianz-Kovarianz-Matrizen und der Sphärizität

Zur Überprüfung der Homoskedastizität zwischen den beiden Faktorstufen des Zwischensubjektfaktors kamen Levene-Tests zum Einsatz (siehe Field, 2017), welche im Rahmen der univariaten varianzanalytischen Überprüfung der Intrusion durch die Art des Nutzertests (Hypothese H1) berechnet wurden und bei allen abhängigen Variablen und Kontrollvariablen Varianzhomogenität bestätigen konnten ($p \geq .05$, siehe Field, 2017). Zur multivariaten Überprüfung der Voraussetzung der Homogenität der Varianz-Kovarianzen wurde ein Boxscher M-Test im Rahmen der Überprüfung der Intrusion durch die Art des Nutzertests (Hypothese H1) berechnet, der jedoch nicht signifikant war ($p \geq .05$, siehe Field, 2017) und dementsprechend die Varianz-Kovarianzen als homogen anzusehen sind. Die Sphärizität zwischen den beiden Faktorstufen des Innersubjektfaktors musste nicht mithilfe von Mauchly-Tests im Rahmen der univariaten varianzanalytischen Überprüfung der Intrusion durch die Art des Nutzertests (Hypothese H1) und der Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H2) berechnet werden, da der Innersubjektfaktor nur zwei Stufen betrug (siehe Field, 2017).

Überprüfung der Voraussetzung der Unabhängigkeit der Fehlerkomponenten

Laut Bortz und Schuster (2011) kann diese Voraussetzung als gegeben angesehen werden, wenn die Versuchspersonen dem Zwischensubjektfaktor randomisiert zugewiesen wurden, was im hier berichteten zweiten Experiment der Fall war.

5.2.4.2 Überprüfung der Validität

Tabelle 5.5. *Deskriptive Daten des zweiten Experiments zur Überprüfung der Intrusion durch die Art des Nutzertests (Hypothese H1) und zur Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H2).*

	Nutzertest mit Rhythmusaufgabe				Nutzertest ohne Rhythmusaufgabe			
	Intuitiv ↓ Fusion 360		Intuitiv ↑ Affinity Des.		Intuitiv ↓ Fusion 360		Intuitiv ↑ Affinity Des.	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
AV_IntuiBeat-S [ms]	132.33	44.28	107.74	41.49	-	-	-	-
AV_CHAI [s]	.24	.26	.44	.34	.38	.33	.57	.28
AV_Effektivität [%]	.52	.43	.76	.15	.66	.41	.78	.15
AV_QUESI [1-5]	1.99	.58	2.85	.73	2.40	.67	2.90	.57
AV_NASA [0-100]	49.12	13.34	38.93	13.71	45.71	14.16	32.84	12.66
AV_SEA [0-220]	95.70	28.26	69.42	30.84	90.26	30.21	62.94	24.09
AV_PhysischeEffizienz [%]	19.06	4.34	36.21	7.41	15.82	3.28	34.72	6.92
AV_ZeitlicheEffizienz [s]	131.61	46.14	217.22	82.35	126.50	53.40	195.33	66.45

Überprüfung einer möglichen Konfundierung durch die Vorerfahrung bei der Nutzung von CUIs

Die Vorerfahrung bei der Nutzung von CUIs (Wertebereich: 0 - 5) beim Nutzertest mit Rhythmusaufgabe (M = 1.31, SD = .23) unterschied sich nicht signifikant von der Vorerfahrung bei der Nutzung von CUIs beim Nutzertest ohne Rhythmusaufgabe (M = 1.32, SD = .15), $t(34)$ = .215, p = .831, d = .07. Laut J. Cohen (1988) kann der Effekt als klein interpretiert werden ($d \leq .5$). Eine konservative post-hoc Analyse der Teststärke mithilfe von G*Power (Faul et al., 2009) mit df = 35 und bei einer angenommenen geringen Effektstärke ($d_{KV\_Vorerfahrung}$ = .07) ergab lediglich eine geringe Teststärke (1 - $\beta_{KV\_Vorerfahrung}$ = .05) im Zuge der Auswertung der Kontrollvariablen KV_Vorerfahrung. Um jedoch eine ausreichend große Power (1 - $\beta \geq .80$) beim festgestellten Effekt erzielen zu können, wären pro Gruppe 3045 Versuchspersonen nötig gewesen, was aufgrund des straffen Zeitplans im Anwenderprojekt, des dortigen Fokus auf qualitative Ergebnisse, des Verständnisses der Meta-Evaluation von IntuiBeat-S als Nebenprodukt und der personellen Einschränkungen im Projekt 3D-GUIde nicht möglich gewesen wäre. Auch bei Annahme einer großen Effektstärke (d = .8), die aufgrund der geringen Anforderungen an CUI-Kenntnisse bei der Stichprobe auch nicht zu erwarten wäre, hätten immerhin noch 26 Versuchspersonen pro Gruppe getestet werden müssen, was im Hinblick auf die Einschränkungen durch das Anwenderprojekt nicht notwendig erschien.

Da die erhobene Vorerfahrung bei der Nutzung von CUIs dennoch einen ungewollten Einfluss auf den Vergleich der beiden Faktorstufen des Zwischensubjektfaktors haben könnte, wurde zusätzlich im Rahmen der Überprüfung der Intrusion durch die Art des Nutzertests (Hypothese H1) mithilfe von Pearson-Produkt-Moment-Korrelationen sichergestellt, dass kein linearer Zusammenhang zwischen der KV_Vorerfahrung und den das Ausmaß an intuitiver Benutzung abbildenden konvergenten Quasi-Außenkriterien bezüglich jeder Kombination des Zwischensubjektfaktors und des Innersubjektfaktors besteht. KV_Vorerfahrung wurde dementsprechend nicht als Kovariate in den folgenden Varianzanalysen berücksichtigt (siehe Döring & Bortz, 2016; Field, 2017).

Überprüfung der Intrusion durch die Art des Nutzertests

Eine multivariate Varianzanalyse mit dem Innersubjektfaktor *unterschiedlich intuitiv benutzbare Software* (d.h. weniger intuitiv benutzbare Software vs. stärker intuitiv benutzbare Software) und dem Zwischensubjektfaktor *Art des Nutzertests* (d.h. Nutzertest mit Rhythmusaufgabe vs. Nutzertest ohne Rhythmusaufgabe) ergab erwartungsgemäß einen signifikanten Haupteffekt der unterschiedlich intuitiv benutzbaren Software bezüglich der, das Ausmaß an intuitiver Benutzung erfassenden, konvergenten Quasi-Außenkriterien AV_QUESI, AV_SEA, AV_NASA, AV_CHAI und AV_Effektivität (H1.A), $F(5,30) = 4.90$, $p = .002$, Pillai-Spur = .45, $\eta_p^2 = .45$. Die Interaktion zwischen der unterschiedlich intuitiv benutzbaren Software und der Art des Nutzertests (H1.C), $F(5,30) = 2.03$, $p = .103$, Pillai-Spur = .25, $\eta_p^2 = .25$, als auch der Haupteffekt des Zwischensubjektfaktors *Art des Nutzertests* (H1.B) waren erwartungsgemäß nicht signifikant, $F(5,30) = .97$, $p = .453$, Pillai-Spur = .139, $\eta_p^2 = .14$. Laut J. Cohen (1988) handelt es sich bei allen Effekten um große Effekte ($\eta_p^2 \geq .14$). Eine konservative post-hoc Analyse der Teststärke mithilfe von G*Power mit $df = 5$ und einer Nullkorrelation innerhalb des Innersubjektfaktors ergab sowohl bei einer angenommenen hohen Effektstärke (d.h. $\eta_p^2 = .25$ für den nicht signifikanten Haupteffekt des Zwischensubjektfaktors) als auch unter Annahme einer geringeren Effektstärke (d.h. $\eta_p^2 = .14$ für den nicht signifikanten Interaktionseffekt beider Faktoren) eine hinreichend große Teststärke (1 - $\beta = 1$). Die deskriptiven Daten der multivariaten Varianzanalyse sind Tabelle 5.5 zu entnehmen.

Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs

Eine multivariate Varianzanalyse mit dem Innersubjektfaktor *unterschiedlich intuitiv benutzbare Software* (d.h. weniger intuitiv benutzbare Software vs. stärker intuitiv benutzbare Software) innerhalb der Faktorstufe *Nutzertest mit Rhythmusaufgabe* des Zwischensubjektfaktors *Art des Nutzertests* ergab erwartungsgemäß einen bedingten signifikanten Haupteffekt bezüglich des durch alle Evaluationsmethoden (d.h. IntuiBeat-S und konvergente Quasi-Außenkriterien) erfassten Ausmaßes an intuitiver Benutzung, $F(6,12) = 3.92$, $p = .021$, Pillai-Spur = .66, $\eta_p^2 = .66$. Laut J. Cohen (1988) handelt es sich es hierbei um einen großen Effekt ($\eta_p^2 \geq .14$)

Die weitere univariate varianzanalytische Auswertung der einzelnen als Quasi-Außenkriterien fungierenden konvergenten Evaluationsmethoden ergab für alle Quasi-Außenkriterien (H2.A bis H2.F) ebenfalls erwartungsgemäß signifikante Effekte (AV_CHAI:

$F(1,17) = 4.80$, $p = .043$, $\eta^2 = .22$; AV_SEA: $F(1,17) = 8.64$, $p = .009$, $\eta^2 = .34$; AV_QUESI: $F(1,17) = 17.17$, $p = .001$, $\eta^2 = .50$; AV_Effektivität: $F(1,17) = 6.24$, $p = .023$, $\eta^2 = .27$; AV_NASA: $F(1,17) = 12.10$, $p = .003$, $\eta^2 = .42$), weswegen die Manipulation als erfolgreich interpretiert werden kann und auch die eingesetzten Evaluationsmethoden empirisch als konvergente Quasi-Außenkriterien für die Meta-Evaluation von IntuiBeat-S bestätigt werden konnten. Bei der univariaten varianzanalytischen Auswertung von IntuiBeat-S konnte auch ein signifikanter Effekt festgestellt werden, $F(1,17) = 4.69$, $p = .045$, $\eta^2 = .22$. Laut J. Cohen (1988) handelt es sich bei allen Effekten um große Effekte ($\eta^2 \geq .14$). Die deskriptiven Daten der abhängigen Variablen und Kontrollvariablen sind Tabelle 5.5 zu entnehmen.

Überprüfung der konvergenten Validität

Tabelle 5.6. *Pearson-Produkt-Moment-Korrelationen zwischen summativen Evaluationsmethoden für intuitive Benutzung (d.h. konvergente Quasi-Außenkriterien) und IntuiBeat-S zur Überprüfung der konvergenten Validität (Hypothese H3) im Zuge des zweiten Experiments.*

	AV_IntuiBeat-S: Intuitiv ↓ Fusion 360	AV_IntuiBeat-S: Intuitiv ↑ Affinity Des.
AV_CHAI	$r(18) = -.60$ $p = .009$	$r(18) = -.65$ $p = .003$
AV_Effektivität	$r(18) = -.59$ $p = .010$	$r(18) = -.58$ $p = .011$
AV_QUESI	$r(18) = -.56$ $p = .016$	$r(18) = -.53$ $p = .023$
AV_NASA	$r(18) = .57$ $p = .013$	$r(18) = .61$ $p = .007$
AV_SEA	$r(18) = .61$ $p = .007$	$r(18) = .68$ $p = .002$

Wie erwartet, zeigten die Pearson-Produkt-Moment-Korrelationen (siehe Tabelle 5.6) innerhalb der Faktorstufe *Nutzertest mit Rhythmusaufgabe* des Zwischensubjektfaktors *Art des Nutzertests*, zwischen den durch IntuiBeat-S erhobenen mittleren Rhythmusabweichungen und den Testwerten der anderen, für die Meta-Evaluation fungierenden konvergenten, Quasi-Außenkriterien zur Überprüfung der konvergenten Validität (Hypothese H3) in beiden Faktorstufen des Innersubjektfaktors *unterschiedlich intuitiv benutzbare Software*, signifikante Korrelationskoeffizienten von $|r| > .50$ ($p < .05$), weswegen hier laut J. Cohen (1988) von hohen linearen Zusammenhängen ($|r| \geq .5$) zwischen den konvergenten Evaluationsmethoden und IntuiBeat-S (H3.A bis H3.E) gesprochen werden kann.

Überprüfung der divergenten Validität

Wie erwartet, zeigten die Pearson-Produkt-Moment-Korrelationen (siehe Tabelle 5.7) innerhalb der Faktorstufe *Nutzertest mit Rhythmusaufgabe* des Zwischensubjektfaktors *Art*

des Nutzertests zwischen den durch IntuiBeat-S erhobenen mittleren Rhythmusabweichungen und den Testwerten der, nicht intuitive Benutzung evaluierenden, divergenten Methoden (d.h. physische Effizienz bei der Systemnutzung und zeitliche Effizienz bei der Handlungsdurchführung) zur Überprüfung der divergenten Validität (Hypothese H4) in beiden Faktorstufen des Innersubjektfaktors *unterschiedlich intuitiv benutzbare Software*, allesamt (H4.A und H4.B) statistisch nicht signifikante lineare Zusammenhänge mit IntuiBeat-S ($p > .05$).

Tabelle 5.7. *Pearson-Produkt-Moment-Korrelationen zwischen summativen Evaluationsmethoden, die nicht intuitive Benutzung messen (d.h. divergente Quasi-Außenkriterien) und IntuiBeat-S zur Überprüfung der divergenten Validität (Hypothese H4) im Zuge des zweiten Experiments.*

	AV_IntuiBeat-S: Intuitiv ↓ Fusion 360	AV_IntuiBeat-S: Intuitiv ↑ Affinity Des.
AV_PhysischeEffizienz	$r(18) = .16$ $p = .522$	$r(18) = -.26$ $p = .303$
AV_ZeitlicheEffizienz	$r(18) = -.08$ $p = .759$	$r(18) = .14$ $p = .568$

5.2.4.3 Überprüfung der zeitlichen Anwendungseffizienz

Wie erwartet, lag die zeitliche Anwendungseffizienz des, mit dem Nutzertest mit Rhythmusaufgabe verbundenen, objektiven Benchmarks *IntuiBeat-S* signifikant höher ($M = 428.17$, $SD = 105.95$) als die zeitliche Anwendungseffizienz des, mit dem Nutzertest ohne Rhythmusaufgabe verbundenen, objektiven Benchmarks *CHAI* ($M = 2324.89$, $SD = 655.99$), $t(34) = 12.11$, $p < .001$, d = 4.04. Es lag laut J. Cohen (1988) ein großer Effekt vor (d $\geq .80$).

5.2.5 Diskussion

Im vorliegenden zweiten Experiment wurde die wissenschaftliche Güte von IntuiBeat-S bezüglich des Gütekriteriums *Validität* im Vergleich mit weiteren, als konvergente und divergente Quasi-Außenkriterien fungierenden, summativen Evaluationsmethoden bei der Nutzung einer weniger intuitiv benutzbaren (Fusion 360, Autodesk, 2017a) und einer stärker intuitiv benutzbaren (Affinity Designer, Serif Inc., 2017) Software empirisch geprüft, unter der Bedingung, dass die konvergenten und divergenten Quasi-Außenkriterien nicht auf Basis eines für die Anwendung von IntuiBeat-S modifizierten Nutzertests mit Rhythmusaufgabe erhoben wurden, sondern auf Basis eines gewöhnlichen Nutzertests ohne zusätzliche Rhythmusnebenaufgabe. Auf diese Weise konnten die Ergebnisse des ersten Experiments sowohl bestätigt als auch eine mögliche Intrusion der Rhythmusaufgabe als Alternativerklärung ausgeschlossen werden.

Anhand der statistischen Tests wurde zunächst überprüft, inwiefern eine Intrusion durch die Art des Nutzertests (d.h. Nutzertest mit Rhythmusaufgabe vs. Nutzertest ohne Rhythmusaufgabe) vorlag (Hypothese H1). Im Anschluss erfolgte die Untersuchung, inwiefern

die Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H2) mithilfe der erhobenen konvergenten Quasi-Außenkriterien und IntuiBeat-S festgestellt werden können. Schließlich wurde überprüft, inwiefern IntuiBeat-S konvergente Validität (Hypothese H3) und divergente Validität (Hypothese H4) unter Berücksichtigung konvergenter und divergenter Quasi-Außenkriterien attestiert werden kann. Darüber hinaus wurde abschließend sichergestellt, dass IntuiBeat-S der in Abschnitt 3.4 als objektiver Benchmark identifizierten CHAI-Methode bezüglich ihrer zeitlichen Anwendungseffizienz überlegen ist (Hypothese H5). Im folgenden Verlauf werden die einzelnen Hypothesen bezüglich der Validität unter Berücksichtigung der festgestellten Ergebnisse diskutiert, nachdem auf eine mögliche Konfundierung durch die Vorerfahrung bei der Nutzung von CUIs eingegangen wurde.

5.2.5.1 Überprüfung der Validität

Überprüfung einer möglichen Konfundierung durch die Vorerfahrung bei der Nutzung von CUIs

Die Vorerfahrung bei der Nutzung von CUIs unterschied sich erwartungsgemäß nicht signifikant zwischen den beiden Faktorstufen des Zwischensubjektfaktors *Art des Nutzertests* (d.h. Nutzertest mit Rhythmusaufgabe vs. Nutzertest ohne Rhythmusaufgabe). Jedoch war die Teststärke aufgrund der geringen beobachteten Effekte und der kleinen Stichprobe gering. Analog zum ersten Experiment zeigte sich kein stärkerer Effekt, obwohl im zweiten Experiment eine differenzierte Operationalisierung der Vorerfahrung anhand beider Dimensionen des konstruierten TFQ (d.h. Häufigkeit und Funktionsumfang) vorgenommen wurde. Wie bereits im Zuge der Diskussion des ersten Experiments erwähnt, könnte eine mögliche Erklärung bereits in der Auswahl der Stichprobe bestehen, da bei den Studierenden keine großen Unterschiede bezüglich der Vorerfahrung existieren. Die praktische Bedeutsamkeit, der in diesem Zusammenhang ermittelten, Effekte bzw. der Erkenntnisgewinn ist demzufolge eher als unbedeutend einzustufen (siehe Döring & Bortz, 2016), was sich auch in den deskriptiven Unterschieden erkennen lässt (siehe Tabelle 5.5).

Da darüber hinaus innerhalb der beiden Faktorstufen des Zwischensubjektfaktors auch keine signifikanten Korrelationen zwischen der Kontrollvariablen KV_Vorerfahrung und den das Ausmaß an intuitiver Benutzung erfassenden konvergenten Evaluationsmethoden festgestellt werden konnte, kann eine Konfundierung durch eine unterschiedliche Vorerfahrung bei der Nutzung von CUIs mit hoher Wahrscheinlichkeit ausgeschlossen werden.

Überprüfung der Intrusion durch die Art des Nutzertests

Wie erwartet konnte bezüglich, der das Ausmaß an intuitiver Benutzung erfassenden, konvergenten Quasi-Außenkriterien ein signifikanter Haupteffekt des Innersubjektfaktors *unterschiedlich intuitiv benutzbare Software* festgestellt werden. Darüber hinaus konnte sowohl kein signifikanter Haupteffekt des Zwischensubjektfaktors *Art des Nutzertests* als auch kein signifikanter Interaktionseffekt beider Faktoren beobachtet werden. Aufgrund der Tatsache, dass die Effektstärke bei beiden Nulleffekten hoch und die Stichprobe relativ klein war, kann eine Intrusion durch die Art des Nutzertests jedoch nicht vollständig

ausgeschlossen werden. Nichtsdestotrotz kann auf Basis der nicht signifikanten Ergebnisse die in diesem Zusammenhang aufgestellten Hypothesen (H1.A, H1.B und H1.C) als bestätigt angesehen werden (siehe Teilabschnitt 5.8). Die Ergebnisse des ersten Experiments wurde also repliziert und eine mögliche Intrusion als Alternativerklärung ist eher unwahrscheinlich.

Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs

Wie erwartet, konnten alle, als konvergente Quasi-Außenkriterien fungierenden, Evaluationsmethoden signifikant ein höheres Ausmaß an intuitiver Benutzung bezüglich der stärker intuitiv benutzbaren Software im Vergleich zur weniger intuitiv benutzbaren Software feststellen. Es ließen sich hierbei durchgängig hohe Effekte beobachten, wodurch sich diese Evaluationsmethoden auch empirisch als konvergente Quasi-Außenkriterien für die Meta-Evaluation von IntuiBeat-S qualifizieren und die in diesem Zusammenhang aufgestellten Hypothesen (H2.A bis H2.F) als bestätigt angesehen werden können (siehe Teilabschnitt 5.8). Der höhere Effekt und das signifikante Ergebnis der AV_Effektivität kann im Vergleich zum ersten Experiment möglicherweise zum Teil darauf zurückgeführt werden, dass die Operationalisierung nun auf Basis der erfolgreich durchgeführten Mausklicks erfolgte und dadurch eine differenzierte Messung der Effektivität möglich war.

Überprüfung der konvergenten Validität von IntuiBeat-S

Wie erwartet zeigten sich hohe signifikante lineare Zusammenhänge (d.h. konvergente Validitätskoeffizienten) zwischen IntuiBeat-S und allen, als Quasi-Außenkriterien fungierenden, konvergenten Evaluationsmethoden innerhalb der Faktorstufe *Nutzertest mit Rhythmusaufgabe*, weswegen alle in diesem Zusammenhang aufgestellten Hypothesen bestätigt werden konnten (H3.A bis H2.E, siehe Teilabschnitt 5.8).

Überprüfung der divergenten Validität von IntuiBeat-S

Wie erwartet zeigten sich nicht signifikant lineare Zusammenhänge (d.h. divergente Validitätskoeffizienten) zwischen IntuiBeat-S und den, nicht intuitive Benutzung evaluierenden, divergenten Methoden innerhalb der Faktorstufe *Nutzertest mit Rhythmusaufgabe*, weswegen alle in diesem Zusammenhang aufgestellten Hypothesen bestätigt werden konnten (H4.A und H4.B, siehe Teilabschnitt 5.8). Dennoch besteht in der Operationalisierung der physischen Effizienz als Anzahl der Klicks bei der Systemnutzung die Gefahr, dass dieses Maß mit Evaluationsmethoden, die die intuitive Benutzung abbilden, bei einer stärker intuitiv benutzbaren Software ungewollt korrelieren könnte. Diese Gefahr konnte mithilfe von Pearson-Produkt-Moment-Korrelationen ($p > .05$) jedoch ausgeschlossen werden.

5.2.5.2 Überprüfung der zeitlichen Anwendungseffizienz

Wie erwartet, lag die die zeitliche Anwendungseffizienz von IntuiBeat-S gegenüber der CHAI-Methode signifikant höher. IntuiBeat-S nahm sogar lediglich ein Fünftel der Zeit

für die Anwendung in Anspruch. Dieser Unterschied lässt sich darauf zurückführen, dass IntuiBeat-S zwar eine explizite Vorbereitungsphase im Vergleich zur CHAI-Methode benötigt (d.h. Baselinemessung schlägt sich mit 60 Sekunden nieder), die Durchführung und Auswertung jedoch wesentlich schneller geht, da diese softwareseitig erfolgt und keine händische Kodierung einzelner Interaktionen erfordert. Die in diesem Zusammenhang aufgestellte Hypothese (H5) kann als bestätigt angesehen werden (siehe Teilabschnitt 5.8).

5.2.6 Schlussfolgerung

Tabelle 5.8. *Übersicht der mithilfe des zweiten Experiments bestätigten Hypothesen im Zuge der Meta-Evaluation von IntuiBeat-S.*

Hypothese	Experiment 2
(H1) Überprüfung der Intrusion durch die Art des Nutzertests:	
- (A) Haupteffekt des Innersubjektfaktors *unterschiedlich intuitiv benutzbare Software*	✓
- (B) Kein Haupteffekt des Zwischensubjektfaktors *Art des Nutzertests*	✓
- (C) Kein Interaktionseffekt des Zwischensubjektfaktors und des Innersubjekfaktors	✓
(H2) Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs:	
- (A) IntuiBeat-S	✓
- (B) SEA	✓
- (C) NASA-RTLX	✓
- (D) QUESI	✓
- (E) Effektivität	✓
- (F) CHAI	✓
(H3) Überprüfung der konvergenten *Validität:*	
- (A) SEA	✓
- (B) NASA-RTLX	✓
- (C) QUESI	✓
- (D) Effektivität	✓
- (E) CHAI	✓
Überprüfung der divergenten Validität:	
- (A) Physische Effizienz	✓
- (B) Zeitliche Effizienz	✓
(H4) Überprüfung der zeitlichen Anwendungseffizienz	✓

Zusammenfassend kann festgehalten werden, dass die wissenschaftliche Güte von IntuiBeat-S als summative Evaluationsmethode für intuitive Benutzung hinsichtlich des Gütekriteriums *Validität* auch unter Berücksichtigung der Art des, als Datengrundlage verwendeten, Nutzertests empirisch bestätigt werden konnte. Konfundierungen durch Unterschiede in der Vorerfahrung bei der Nutzung von CUIs können mit hoher Wahrscheinlichkeit ausgeschlossen werden und es lagen durchgehend hohe Effekt-, sowie Teststärken vor. Darüber hinaus konnte IntuiBeat-S auch ein für das Projekt 3D-GUIde wichtiger Teilaspekt praktischer Güte nachgewiesen werden, indem IntuiBeat-S gegenüber der als objektiver Benchmark fungierenden CHAI-Methode eine höhere zeitliche Anwendungseffizienz zeigte. Diese Ergebnisse replizieren die Ergebnisse des ersten Experiments, während sie eine Intrusion durch die Art des als Datengrundlage verwendeten Nutzertests ausschlossen.

Jedoch lassen sich noch die, bei beiden Experimenten als Analogstichprobe eingesetzten, Studierenden der Medienkommunikation und der Mensch-Computer-Systeme (d.h. mangelnde Generalisierbarkeit von IntuiBeat-S aufgrund der verwendeten Stichprobe), die nicht Kernanwender von CUIs sind (d.h. besitzen dementsprechend lediglich geringe Vorerfahrung bei der Nutzung von CUIs), und das in beiden Experimenten verwendete Pedal eines bestimmten Herstellers (d.h. mangelnde Generalisierbarkeit von IntuiBeat-S aufgrund eines bestimmten Aufzeichnungsgeräts), dass nicht für jedermann zugänglich ist, als Limitationen der beiden durchgeführten Experimente anführen. Für einen problemlosen Einsatz von IntuiBeat-S im Projekt 3D-GUIde müssen diese beiden offenen Punkte noch geklärt werden, damit diverse Evaluatoren mit unterschiedlichen Arten von Pedalen die Systemnutzung von verschiedenen Expertennutzern in verschiedenen Domänen (d.h. CUIs und Aufgaben, die auch in ingenieurmäßigen Fächern Anwendung finden) summativ beurteilen können. Eine konzeptuelle Replikation der ersten beiden Experimente mit (1) zwei anderen unterschiedlich intuitiv benutzbaren CUIs, (2) der Berücksichtigung einer anderen Stichprobe (d.h. Studierende aus Studiengängen und verschiedenen Bildungseinrichtungen, in denen die Nutzung von CUIs einen höheren Schwerpunkt bildet und die Studierenden entsprechend über mehr Vorerfahrung verfügen) und eines (3) quelloffenen, jedermann zugänglichen Pedals sollte alle offenen Fragen klären können, die damit in Zusammenhang stehen.

5.3 Experiment 3

Das in diesem Abschnitt vorgestellte dritte Experiment verfolgte das Ziel zu überprüfen, inwiefern IntuiBeat-S auch wissenschaftliche Güte bezüglich der formalen Hauptgütekriterien *Objektivität*, *Reliabilität* und *Validität* im Vergleich mit weiteren, als konvergente, sowie divergente Quasi-Außenkriterien fungierenden, summativen Evaluationsmethoden bei der Nutzung eines weniger intuitiv benutzbaren (Rhinoceros 3D, Robert McNeel & Associates, 2017) und eines stärker intuitiv benutzbaren (Fusion 360, Autodesk, 2017a) CUI aufweist, wenn für die Meta-Evaluation eine heterogenere Stichprobe (d.h. Studierende diverser MINT-Fächer aus verschiedenen Institutionen, die über mehr Vorerfahrung bei der Nutzung von CUIs verfügen) als bei den beiden vorherigen Experimenten (d.h. Studierende der Mensch-Computer-Systeme und der Medienkommunikation der Universität Würzburg) und eine andere Aufzeichnungshardware (d.h. quelloffenes Fußpedal anstelle

eines Fußpedals eines bestimmten Herstellers) verwendet würden. Es wurde dementsprechend erneut die erste Forschungsfrage dieser Arbeit durch das Experiment betrachtet.

5.3.1 Überprüfung der Objektivität von IntuiBeat-S

Wie bereits beim ersten Experiment (siehe Abschnitt 5.1) und dem Folgeexperiment wurde die Objektivität nicht weiter empirisch untersucht, da aufgrund der hohen Standardisierung von IntuiBeat-S dies nicht zwingend notwendig erschien. Das neu entwickelte Pedal „Taktschuh“ sollte darauf auch keinen Einfluss ausüben. Des Weiteren übernahmen die Durchführung dieses und der beiden vorigen Experimente verschiedene Evaluatoren unter Anleitung des Versuchsleiters, was zusätzlich die Objektivität von IntuiBeat-S demonstrierte, so wie es bereits in der Einleitung dieses Kapitels und in Teilabschnitt 5.1.3 angesprochen wurde.

5.3.2 Überprüfung der Reliabilität von IntuiBeat-S

Da im vorliegenden Experiment die Rhythmusaufzeichnung mithilfe anderer Hardware erfolgte, erschien eine erneute Überprüfung der Reliabilität von IntuiBeat-S sinnvoll. Wie bereits beim ersten Experiment (siehe Abschnitt 5.1) wurde als methodischer Zugang zur Beurteilung der Reliabilität die Testhalbierungs-Reliabilität gewählt, welche konkret mit dem Guttman Reliabilitätskoeffizienten operationalisiert wurde. Für weitere Details bezüglich der Überprüfung der Reliabilität von IntuiBeat-S sei an dieser Stelle auf die Hypothesen des ersten Experiments verwiesen (siehe Teilabschnitt 5.1.3).

5.3.3 Überprüfung der Validität von IntuiBeat-S

Wie bereits beim ersten Experiment (siehe Abschnitt 5.1) wurde als methodischer Zugang zur Beurteilung der Validität die Konstruktvalidität im Sinne eines struktursuchenden Vorgehens durch Sicherstellung der konvergenten und divergenten Validität gewählt. Für weitere Details bezüglich der Überprüfung der Validität von IntuiBeat-S sei an dieser Stelle auf das erste Experiment verwiesen (siehe Teilabschnitt 5.1.3). Die an dieser Stelle aufgeführten Hypothesen wurden für das dritte Experiment ohne Änderungen übernommen.

5.3.4 Methode

5.3.4.1 Teilnehmer

Für das dritte Experiment wurden 11 Versuchspersonen über Mailverteiler der Fachschaft für Luft- und Raumfahrttechnik der Universität Stuttgart, sowie direkt aus dem persönlichen Bekanntenkreis des Evaluators rekrutiert. Da für die Meta-Evaluation von IntuiBeat-S nur vollständige Datensätze berücksichtigt werden konnten, musste eine Versuchsperson für die spätere Datenauswertung direkt ausgeschlossen werden, da die Studie wegen

technischer Probleme bei der Videoaufzeichnung abgebrochen werden musste. Demzufolge konnten für die Meta-Evaluation von IntuiBeat-S 10 Versuchspersonen berücksichtigt werden, welche alle rechtsfüßig (d.h. der rechte Fuß stellte den dominanten Fuß dar und wurde für die Rhythmuseingabe genutzt) waren. Die Stichprobe setzten sich dabei aus einer Frau und neun Männern zusammen. Das Durchschnittsalter betrug 24.30 Jahre ($SD = 2.91$). Vier Versuchspersonen (40 %) studierten an der Universität Stuttgart Luft- und Raumfahrttechnik (B.Sc.), eine Versuchsperson (10 %) studierte an der Universität Stuttgart Maschinenbau (B.Sc.), eine Versuchsperson (10 %) studierte an der Universität Stuttgart Erneuerbare Energien (B.Sc.), eine Versuchsperson (10 %) studierte an der Universität Stuttgart Photonic Engineering (M.Sc.), eine Versuchsperson (10 %) studierte Maschinenbau (B.Sc.) an der Dualen Hochschule Baden-Württemberg Stuttgart und eine Versuchsperson (10 %) studierte Maschinenbau (B.Sc.) an der Hochschule Pforzheim. Eine Versuchsperson (10 %) übte bereits den Beruf des technischen Produktdesigners aus und hatte demnach keinen Studentenstatus mehr. Bei der Rekrutierung wurden alle Versuchspersonen über eine gesonderte Mail darauf hingewiesen, für den Versuch flache Sportschuhe zu tragen, um eine möglichst problemlose Rhythmuseingabe über das USB-Fußpedal zu ermöglichen. Die Vorerfahrung der Versuchspersonen bezüglich der Nutzung von CUIs betrug im Durchschnitt 4.55 ($SD = 1.14$) bei einem Maximum von 6 und lagen damit erwartungsgemäß im oberen Bereich. Alle Versuchspersonen bekamen als Gegenleistung für die Teilnahme zwei Tafeln Schokolade und gaben an freiwillig am Experiment teilzunehmen.

5.3.4.2 Versuchsdesign

Für die Beantwortung der ersten Forschungsfrage unter Berücksichtigung einer heterogeneren Stichprobe im Vergleich zu den vorherigen Experimenten und zum Nachweis der Robustheit von IntuiBeat-S wurde ein einfaktorielles experimentelles Within-Subjects Design genutzt. Die unabhängige Variable war die *unterschiedliche intuitiv benutzbare Software* mit den Ausprägungen: weniger intuitiv benutzbare Software vs. stärker intuitiv benutzbare Software. Die abhängige Variable stellte das *Ausmaß an intuitiver Benutzung* dar.

5.3.4.3 Versuchsmaterialien und Maße

Unterschiedlich intuitiv benutzbare Software

Zur Operationalisierung der unabhängigen Variable *unterschiedlich intuitiv benutzbare Software* wurden auf Basis einer qualitativen Experteneinschätzung ($N_{Experte}$= 5; Vorgehen: siehe entsprechenden Absatz innerhalb des Teilabschnitts 5.1.4.3) das 3D-CUI *Rhinoceros 3D* von Robert McNeel & Associates (2017) als wenig intuitiv gestaltete Software (d.h. erwartete hohe Diversität und Komplexität der Systemnutzung, da das Experten-Tool für die Abbildung des Gestaltungs-, Entwicklungs- und Fertigungsprozess auf NURBS, also Non-Uniform Rational B-Splines, auf mathematische Kurven anstelle von Polygonen zurückgreift. Aus Polygonen lässt sich direkt ein 3D-Modell erschaffen, wohingegen bei den viel genaueren NURBS noch als Zwischenschritt manuell Flächen erstellt

werden müssen, um damit letztendlich ein 3D-Modell aufbauen zu können, siehe Chopine, 2012) und das 3D-CUI *Fusion 360* von Autodesk (2017a) als stärker intuitiv gestaltete Software (d.h. erwartete geringe Diversität und Komplexität, da das Experten-Tool zwar den Anspruch erhebt, den vollständigen Gestaltungs-, Entwicklungs- und Fertigungsprozess abzubilden, aber für die 3D-Modellierung direkt auf Polygone anstelle von NURBS zurückgegriffen werden kann, siehe Chopine, 2012) für die Meta-Evaluation von IntuiBeat-S ausgewählt (siehe Abbildung 5.10). Fusion 360 wurde im Vergleich zu den Vorgängerexperimenten auf Basis von Evalint (Scholz, 2006) außerdem als stärker intuitiv eingestuft, da die Stichprobe sich nun aus studentischen Expertennutzern zusammensetzte. Um diese miteinander vergleichen zu können, wurden von gleichen Experten verschiedene experimentelle Aufgaben mit der gleichen Zielsetzung wie im ersten Experiment (siehe Abschnitt 5.1) gewählt, die in ein fiktives Szenario eingebettet waren, in der es um die Entwicklung eines Elektromotors für ein neues ferngesteuertes Modellauto ging. Die Aufgaben wurden in einer randomisierten Reihenfolge bearbeitet.

Aufgrund der Tatsache, dass für dieses Experiment weniger Studierende im Vergleich zu den vorigen beiden Experimenten verfügbar waren, da sie zur Gewährleistung hoher Heterogenität nicht zentral an einer Universität rekrutiert und einen stärkeren ingenieurwissenschaftlichen Hintergrund als die bisherigen Stichproben aufweisen sollten, und bisherige Experimente keine Konfundierung durch die Vorerfahrung zeigen konnten, wurde sich für ein Within-Subjects Design anstelle eines Between-Subjects Design entschieden. Um, wegen der inhaltlichen Ähnlichkeit der experimentellen Aufgaben und des Messwiederholungsdesigns, die Ergebnisse nicht durch etwaige Übungseffekte zu konfundieren, wurde sich ferner entschieden, die Darbietungsreihenfolge der CUIs für jede Versuchsperson zufällig erfolgen zu lassen und die Reihenfolgen über alle Versuchspersonen hinweg auszubalancieren.

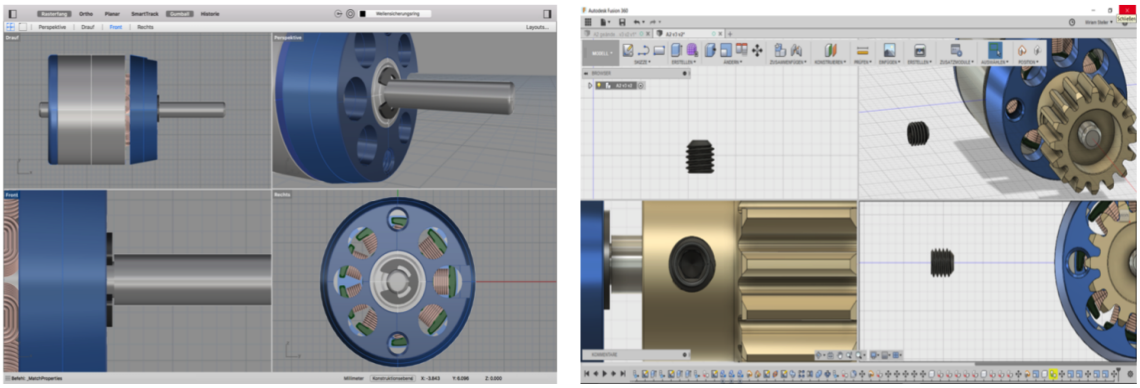


Abbildung 5.10. Die getesteten 3D-CUIs *Rhinoceros 3D* (links, Robert McNeel & Associates, 2017) und *Fusion 360* (rechts, Autodesk, 2017a), welche die beiden Faktorstufen (d.h. Rhinoceros 3D: weniger intuitiv gestaltete Software; Fusion 360: stärker intuitiv gestaltete Software) der unabhängigen Variable *unterschiedlich intuitiv gestaltete Software* repräsentierten.

Alle Versuchspersonen bearbeiteten dementsprechend die drei experimentellen Aufgaben nacheinander mit beiden Faktorstufen: *Fusion 360* (siehe Anhang B.3.1.1) und *Rhinoceros 3D* (siehe Anhang B.3.1.2). Bei den gewählten Aufgaben wurde, wie bei den vorigen Experimenten, darauf geachtet, dass die von LaViola et al. (2017) beschriebenen domä-

nenübergreifenden, grundlegenden Nutzeraufgaben abgedeckt waren (siehe Absatz 5.1.4.3 des ersten Experiments), um möglichst viele qualitative Informationen für die Entwicklung von Interaktionspatterns ableiten zu können. Die erste Aufgabe erforderte es, dass die Versuchspersonen das Ritzel der Antriebseinheit des Elektromotors im Raum drehen mussten (erste Aufgabe). Die nächste Aufgabe sah es vor, dass die Versuchspersonen die Antriebswelle und den Stator zusammensetzen mussten (zweite Aufgabe). Die letzte Aufgabe forderte von den Versuchspersonen, den Gewindestift der Antriebseinheit in Breite und Länge zu skalieren (dritte Aufgabe).

Summative Evaluation intuitiver Benutzung mit IntuiBeat-S

Wie bereits im ersten und zweiten Experiment, stellte das Ausmaß an intuitiver Benutzung auch eine abhängige Variable im dritten Experiment dar. Als zentrale objektive Methode wurde ebenfalls *IntuiBeat-S* zur Operationalisierung dieser Variable genutzt (d.h. AV_IntuiBeat-S), da deren wissenschaftliche Güte im Rahmen des dritten Experiments unter Berücksichtigung einer heterogeneren Stichprobe und anderer Hardware untersucht werden sollte, um die generelle Robustheit der Methode auch in einer heterogenen Umgebung zu demonstrieren. Die Vorgehensweise und die Testumgebung für die summative Evaluation intuitiver Benutzung mithilfe von IntuiBeat-S blieb im dritten Experiment dementsprechend gleich und ist somit mit der in Teilabschnitt 5.1.4.3 enthaltenen Beschreibung des ersten Experiments identisch.

Funktionsweise des neuen USB-Fußpedals „Taktschuh“

Der einzige Unterschied bezüglich der Anwendung von IntuiBeat-S bestand damit in der, im Rahmen des dritten Experiments eingesetzten, Hardware. Im Gegensatz zu den ersten beiden Experimenten wurde im dritten Experiment ein USB-Fußpedal verwendet, dass eine Eingabe über die Ferse erlaubte. Es wurde sich für ein solches Fußpedal entschieden, um zum einen zu demonstrieren, dass es unerheblich ist, wie die Versuchspersonen den Rhythmus eingeben (d.h. Eingabe durch den Fußballen und Belastung der vorderen Schienbeinmuskulatur bzw. Eingabe durch die Ferse und Belastung der hinteren Wadenmuskulatur) und damit zum anderen gleichzeitig ein erprobtes, für alle potentiellen Anwender verfügbares USB-Fußpedal bereitzustellen.

Es wurde sich dementsprechend gegen ein bestimmtes Fabrikat eines bestimmten Herstellers entschieden, so wie es in den ersten beiden Experimenten der Fall war, sondern stattdessen ein eigenes USB-Fußpedal namens „Taktschuh“ am Lehrstuhl für Psychologische Ergonomie konstruiert und das Gehäuse mit dem 3D-Drucker *ZMorph VX Multitool 3D Printer* gedruckt (siehe Abbildung 5.11). Bei der Funktionalität des Taktschuhs und der technischen Umsetzung wurde sich an den zuvor verwendeten USB-Fußpedalen orientiert, die nichts anderes als externe Tastaturen sind, die bei Betätigung des Fußschalters ein Zeichen ausgeben (z.B. „0“). Um eine solche Funktionalität selbst implementieren zu können, konnte softwareseitig auf das in der Programmiersprache C geschriebene, unter der GNU General Public License Version 2 (GPL) zur Verfügung gestellte Projekt EasyLogger (Objective Development Software GmbH, 2012) zurückgegriffen und der Code leicht abgeändert werden, um bei Betätigung eines im Taktschuh verklebten Interlink

Drucksensors eine „0“ an die IntuiBeat-Software zu übermitteln, und so eine Eingabe identifizieren zu können. Die softwareseitige und die ergänzende hardwareseitige Implementierung (d.h. Konstruktion des Taktschuhs, Drucken des Taktschuhs, Anordnung der elektronischen Bauteile, etc.) wurde dankenswerterweise von Frank Seyfarth (Technischer Mitarbeiter des Lehrstuhls für Psychologische Ergonomie) übernommen. Für die genaue Liste der elektronischen Bauteile, den Schaltplan dieser Bauteile, den angepassten C-Code und die Konfiguration der Fuse-Bits des Mikrocontrollers sei auf das GitLab-Repository „Taktschuh“ (Reinhardt, 2019) verwiesen. Dieses Repository enthält auch weiterführende Informationen und die für den 3D-Druck benötigten Dateien.

Abbildung 5.11. Das im Rahmen des dritten Experiments eingesetzte USB-Fußpedal „Taktschuh“.

Summative Evaluation intuitiver Benutzung mit anderen Methoden

Wie bereits im ersten und zweiten Experiment wurde, zur Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H1) und zur Überprüfung der konvergenten Validität (Hypothese H2), die abhängige Variable *Ausmaß an intuitiver Benutzung* noch anhand von drei subjektiven Methoden und zwei objektiven Methoden operationalisiert, die als konvergente Quasi-Außenkriterien genutzt wurden. Es kamen hierzu die gleichen Operationalisierungen und die gleiche Administrationsweise wie im zweiten Experiment zum Einsatz, weswegen an dieser Stelle für weitere Informationen auf den entsprechenden Absatz des Teilabschnitts 5.2.3.3 des zweiten Experiments verwiesen wird. Das zweite Experiment operationalisierte die abhängige Variable AV_Effektivität im Vergleich zum ersten Experiment, als Anteil erfolgreich durchgeführter Mausklicks an der Gesamtzahl aller getätigten Mausklicks feiner und erlaubte somit eine präzisere Erfassung des Merkmals.

Summative Evaluation von physischer und zeitlicher Effizienz

Wie bereits im ersten und zweiten Experiment wurden zur Überprüfung der divergenten Validität (Hypothese H3) noch zwei weitere objektive Methoden als divergente Quasi-Außenkriterien genutzt. Es kamen hierzu die gleichen Operationalisierungen und die gleiche Administrationsweise wie im ersten und zweiten Experiment zum Einsatz, weswegen an

dieser Stelle für weitere Informationen auf den entsprechenden Absatz des Teilabschnitts 5.1.4.3 des ersten Experiments verwiesen wird.

Kontrollvariablen und demographische Variablen

Im Rahmen des dritten Experiments wurden neben den vorgestellten Evaluationsmethoden noch demographische Daten erhoben. Wie bereits im zweiten Experiment wurden diese in Papierform administriert. Die Art der soziodemografischen Daten (d.h. Alter, Geschlecht und Studiengang) war, wie in beiden Vorgängerexperimenten, identisch (siehe Absatz „Vorerfahrung bei der Nutzung von CUIs“ des Teilabschnitts 5.1.4.3 des ersten Experiments).

Im Gegensatz zu den vorigen Experimenten, stellte die *Vorerfahrung bei der Nutzung von CUIs* lediglich eine weitere demographische Information dar, da in bisherigen Experimenten keine Konfundierung durch die Vorerfahrung festgestellt werden konnte und zusätzlich ein Messwiederholungsdesign zum Einsatz kam (d.h. beide CUIs wurden immer von einer Versuchsperson mit einer bestimmten Vorerfahrung benutzt und damit dieser Aspekt durch das Messwiederholungsdesign kontrolliert). Nichtsdestotrotz wurde die *Vorerfahrung bei der Nutzung von CUIs* mit einem TFQ erhoben, der wie im Experiment zuvor lediglich die beiden getesteten CUIs „Fusion 360“ und „Rhinoceros 3D“, sowie eine Wildcard „Andere CAD-Software?“ als Items enthielt, um mithilfe der Wildcard die Eingabe eines beliebigen bekannten CUI zu gestatten. Wie in den Experimenten zuvor, wurde sich für die Verwendung des Begriffs „CAD“ anstelle von „CUI“ entschieden, da dieser Begriff in der genutzten Stichprobe wahrscheinlich geläufiger ist und damit keiner Einführung bedarf. Versuchspersonen beurteilten die Items des TFQ dann bezüglich deren Nutzungshäufigkeit (Wertebereich: 0 - 6) und des genutzten Funktionsumfangs (Wertebereich: 0 - 4). Der Gesamtmittelwert über beide Mittelwerte beider Dimensionen stellte dann die Operationalisierung der *Vorerfahrung bei der Nutzung von CUIs* dar (Wertebereich: 0 - 5) und wurde, wie im zweiten Experiment berechnet (siehe entsprechender Absatz des Teilabschnitts 5.2.3.3 des zweiten Experiments). Da kein einheitlicher TFQ existiert und dieser jeweils für die getesteten Softwareanwendungen erstellt werden musste, ist der in diesem Experiment genutzte TFQ vollständig in Anhang B.3.2 dieser Arbeit zu finden.

Apparatur

Die im Rahmen des dritten Experiments verwendete Apparatur unterschied sich in einigen Punkten von der im ersten Experiment genutzten Apparatur. Da das dritte Experiment in den Räumlichkeiten der Hochschule der Medien in Stuttgart durchgeführt wurde, musste generell auf andere Hardware und eine andere räumliche Anordnung der Apparatur zurückgegriffen werden. Wie bereits im ersten Experiment wurde Morae von TechSmith (2004) für die Bildschirmaufzeichnung und die Aufzeichnung von Mauswegen verwendet (siehe Teilabschnitt 5.1), weswegen die Apparatur aus Versuchspersonen-PC, Evaluatoren-PC und Analyse-PC bestand, die horizontal im Raum angeordnet war (siehe Abbildung 5.12). Wie bereits im zweiten Experiment kam anstelle eines Drehstuhls ein gewöhnlicher Stuhl als Sitzgelegenheit für die Versuchsperson zum Einsatz.

Abbildung 5.12. Apparatur des dritten Experiments, bestehend aus Versuchspersonen-PC (linke Seite der Trennwand), Analyse-PC (rechte Seite der Trennwand, linker PC) und Evaluatoren-PC (rechte Seite der Trennwand, rechter PC).

Als Versuchspersonen-PC kam ein 17 Zoll Notebook G752V von Asus (Windows 10 Home, 2.6 GHz Quad-Core Intel i7, 8 GB Arbeitsspeicher, 6 GB NVIDIA GeForce GTX 1060 Grafikkarte, Auflösung 1920 x 1080 Pixel) zum Einsatz, auf dem vom Evaluator die beiden getesteten CUIs installiert wurden. Außerdem wurde für die spätere Aufzeichnung die Software Morae Recorder installiert. Aufgrund der Größe des eingebauten Displays wurde für den Versuchspersonen-PC kein dedizierter externer Monitor verwendet. Im Gegensatz zu den vorherigen Experimenten wurde außerdem keine externe Tastatur verwendet, sondern auf die eingebaute Tastatur des Laptops zurückgegriffen. Als Maus kam eine kabelgebundene Maus B100 von Logitech zum Einsatz. Der Abstand zwischen den Versuchspersonen und dem Laptopbildschirm betrug circa 50 cm. Der Versuchspersonen-PC war von den anderen beiden PCs durch eine Trennwand abgeschirmt, um den Arbeitsplatz der Versuchsperson von dem des Evaluators zur Vermeidung etwaiger Ablenkungen zu trennen.

Aufgrund der Verwendung von Morae als Aufzeichnungssoftware kam wie beim ersten Experiment ein dedizierter Evaluatoren-PC in Form eines 15 Zoll Notebooks ThinkPad T460s Signature Edition von Lenovo (Windows 10 Home, 2.6 GHz Dual-Core Intel i7, 12 GB Arbeitsspeicher, Intel HD 520 Grafikkarte, Auflösung 1920 x 1080 Pixel) zum Einsatz, auf dem die Software Morae Observer installiert wurde. Im Vergleich zum ersten Experiment wurde aufgrund des relativ kleinen Arbeitsplatzes des Evaluators auf einen externen Monitor und externe Eingabegeräte verzichtet (siehe Abbildung 5.12). Der Abstand zwischen dem Evaluator und dem Evaluatoren-PC betrug circa 50 cm. Als letzter Teil der Apparatur fungierte schließlich ein 13 Zoll MacBook Air (MacOS High Sierra, 1.6 GHz Dual-Core Intel i5, 8 GB Arbeitsspeicher, 1536 MB Intel HD Graphics 6000 Grafikkarte, Auflösung 1440 x 900 Pixel) als Analyse-PC. Im Gegensatz zum ersten Experiment, wurde die Software Morae Manager nicht auf dem Analyse-PC, sondern auf dem Evaluatoren-PC installiert (siehe Absatz 5.1.4.3). Auf dem Analyse-PC wurde dementsprechend nur die IntuiBeat-Software ausgeführt und das entwickelte USB-Fußpedal „Taktschuh“ per USB 2.0 angeschlossen. Die Aufzeichnungsdateien wurden zwischen den Systemen (d.h. Versuchspersonen-PC, Evaluatoren-PC und Analyse-PC) über ein gemeinsames Netzlauf-

werk ausgetauscht. Da der Analyse-PC nur während des Erhebungszeitraums zur Verfügung stand, wurde für die retrospektive Auswertung der CHAI-Methode der im ersten Experiment genutzte Analyse-PC verwendet, auf dem der VLC Mediaplayer (VideoLAN, 2017) in Version 2.2.8 bereits installiert war (siehe Absatz „Apparatur“ innerhalb des Teilabschnitts 5.1.4.3 des ersten Experiments).

5.3.4.4 Versuchsdurchführung

Im Gegensatz zu den beiden vorigen Experimenten wurden die Versuchspersonen nicht in Würzburg einheitlich über das Probandensystem des Instituts Mensch-Computer-Medien, sondern in Stuttgart über den E-Mail-Verteiler der Fachschaft Luft- und Raumfahrttechnik der Universität Stuttgart (siehe Anhang A.2), sowie dem persönlichen Bekanntenkreis der Evaluatorin rekrutiert. Alle über diese Kanäle angebotenen Termine waren wie bei den beiden Experimenten zuvor Einzeltermine, da auch immer nur eine Person gleichzeitig getestet wurde.

Das dritte Experiment wurde in einem großen, reizabgeschirmten, stimulusarmen Labor („UX-Lab“) der Information Experience and Design Research Group der Hochschule der Medien Stuttgart durchgeführt, welches im Vergleich zu den beiden vorherigen Experimenten das größte Labor für die Meta-Evaluation von IntuiBeat-S darstellte. Der Geräuschpegel von außen war über den gesamten Erhebungszeitraum konstant niedrig. Bevor die Versuchspersonen das Testlabor betreten durften, wurden sie vom Evaluator mündlich aufgefordert, ihre Mobiltelefone in den Flugmodus zu schalten, um eine Ablenkung durch diese während der Untersuchung zu vermeiden. Die eigentliche Datenerhebung der Studie unterteilte sich in drei Phasen (d.h. Vorphase, Hauptphase, Abschlussphase), die aufgrund des Messwiederholungsdesigns, der Komplexität der Experimentalaufgaben und des retrospektiven Interviews circa 108 Minuten in Anspruch nahmen. Wie bereits beim zweiten Experiment soll im Folgenden nur auf die Unterschiede in den einzelnen Phasen der Versuchsdurchführung im Vergleich zum ersten Experiment eingegangen werden und für weitere Details bezüglich der Gemeinsamkeiten auf Teilabschnitt 5.1.4.4 des ersten Experiments verwiesen werden.

Vorphase (3 Minuten)

Wie bereits im zweiten Experiment wurde auch beim dritten Experiment die Auswirkung der rhythmischen Wahrnehmung auf die wissenschaftliche Güte von IntuiBeat-S nicht untersucht. Im Gegensatz zu den beiden vorherigen Experimenten wurden Kontrollvariablen und demographische Daten außerdem erst in der Abschlussphase erfragt, weswegen sich die Vorphase lediglich auf eine kurze standardisierte mündliche Begrüßung beschränkte, die das Ausfüllen einer schriftlichen Einverständniserklärung umfasste (siehe Anhang A.4).

Hauptphase (95 Minuten)

Wie beim ersten Experiment (siehe Teilabschnitt 5.1.4.4) erhielten die Versuchspersonen zu Beginn der Hauptphase zunächst eine standardisierte mündliche Instruktion. Im

Rahmen dieser Instruktion wurde den Versuchspersonen zunächst der genaue Ablauf des Nutzertests mitgeteilt und ihnen dabei erläutert, dass das Ziel eines solchen Tests darin bestehe, sowohl Probleme des Systems zu identifizieren als auch das Ausmaß an intuitiver Benutzung der CUIs zu bewerten. An dieser Stelle teilte der Evaluator den Versuchspersonen ebenfalls mit, dass die intuitive Benutzung neben anderen Methoden (z.B. Fragebögen) auch mithilfe einer Rhythmuszweitaufgabe im Rahmen der Hauptphase gemessen werden soll. Im Anschluss wurde wie im ersten Experiment eine Baseline-Messung für jede Versuchsperson durchgeführt und eine Einführung in IntuiBeat-S gegeben, so wie sie im Rahmen des ersten Experiments im entsprechenden Absatz des Teilabschnitts 5.1.4.3 ausführlich beschrieben wurde. Die Einführung und die Baseline-Messung nahmen circa fünf Minuten in Anspruch.

Daraufhin stellte der Evaluator den Versuchsteilnehmern standardisiert mündlich ein kurzes, fiktives Konstruktionsszenario vor, welches als Rahmenhandlung für die mit den CUIs zu erledigenden Testaufgaben dienen sollte. Die darauffolgende Aufgabenbearbeitung gestaltete sich identisch zum ersten Experiment, weswegen für genauere Informationen auf den entsprechenden Absatz 5.1.4.4 des ersten Experiments verwiesen werden soll. Der einzige Unterschied zum ersten Experiment bestand neben den anderen getesteten CUIs und damit verbundenen Testaufgaben also darin, dass diese CUIs immer allen Versuchspersonen randomisiert dargeboten wurden (d.h. Innersubjektfaktor), weswegen die Darbietungsreihenfolge der CUIs über alle Versuchspersonen hinweg ausbalanciert und nach jedem bearbeiteten CUI mit den Versuchspersonen ein retrospektives Interview geführt wurde. Da es sich wegen der Expertenzielgruppe um komplexere Aufgaben handelte, wurde für jede Aufgabe innerhalb der jeweiligen Software eine gesonderte 3D-Szene des Elektromotorszenarios auf dem Versuchspersonen-PC (siehe Teilabschnitt 5.1.4.3) vom Evaluator geladen (Dateien „FusionExperiment3A1.f3d“, „FusionExperiment3A2.f3d“, „FusionExperiment3A3.f3d“ für Fusion 360; Dateien „RhinoExperiment3A1.3dm“, „RhinoExperiment3A2.3dm“, „RhinoExperiment3A3.3dm“ für Rhinoceros 3D).

Dieses retrospektive Interview war im Vergleich zu den Interviews, die im Rahmen der ersten beiden Experimente geführt wurden, etwas länger, da die Aufgaben etwas komplexer waren und zudem auf die Erfassung von Aussagen der Versuchsteilnehmer bezüglich ihren positiver Erfahrung beim Umgang mit Technik (siehe Zeiner, Laib, Schippert, & Burmester, 2016) geachtet wurde. Außerdem wurde bei jeder Versuchsperson eine kleine Pause (circa 10 Minuten) zwischen den CUIs eingelegt, um Übertragungseffekte zwischen den beiden CUIs weiter zu minimieren. Das Abbruchkriterium von fünf Minuten, welches bei den beiden vorherigen Experimenten als maximale Bearbeitungsdauer einer Arbeitsaufgabe zum Einsatz kam, wurde in diesem Experiment ebenfalls beibehalten. Die Hauptphase des Experiments erstreckte sich damit auf 45 Minuten für jedes CUI und fünf Minuten für die Einführung in IntuiBeat-S, sowie die damit verbundene Baseline-Messung (Pausenzeit nicht mit eingerechnet).

Abschlussphase (10 Minuten)

Abschließend wurden von den Versuchspersonen die Kontrollvariablen und demographischen Variablen in Papierform abgefragt, so wie dies in der Vorphase der vorigen beiden

Experimente bereits ausführlich beschrieben wurde (siehe Teilabschnitt 5.1.4.4 des ersten und Teilabschnitt 5.2.3.4 des zweiten Experiments). Wie bereits bei den vorherigen beiden Experimenten klärte der Evaluator die Versuchsteilnehmer im Anschluss standardisiert mündlich über die wahre Untersuchungsabsicht (d.h. Debriefing mit Informationen zu Studienzielen und Hypothesen) auf. Der Evaluator erkundigte sich außerdem persönlich nach dem Befinden jedes Teilnehmers und versuchte dabei mögliche Unklarheiten bezüglich des Experiments zu beseitigen. Darüber hinaus wurde jede Versuchsperson darauf hingewiesen, dass es für den weiteren Erfolg der Untersuchung unabdingbar sei, dass gegenüber potentiellen Versuchsteilnehmern nicht über die Inhalte der Studie gesprochen werde. Im Anschluss notierte sich der Evaluator den Namen der jeweiligen Versuchsperson auf einer Teilnehmerliste, um ihr die Versuchspersonenstunde verbuchen zu können. Um die Anonymität der Teilnehmer zu wahren, wurde diese getrennt von den erhobenen Daten aufbewahrt und lag lediglich physisch vor. Des Weiteren bedankte sich der Evaluator für die Teilnahme und verabschiedete die Versuchsperson. Der Evaluator setzte am Ende des Experiments den Versuchspersonen-PC (d.h. Schließen des getesteten CUI) und den Analyse-PC (d.h. Schließen und erneutes Öffnen der IntuiBeat-Software) in den Ausgangszustand zurück und kontrollierte die aufgezeichneten Daten auf Vollständigkeit.

5.3.4.5 Statistische Auswertung

Die statistische Datenauswertung erfolgte mittels IBM SPSS Statistics 25 für macOS. Zunächst wurde vor der eigentlichen Datenauswertung geprüft, ob die Daten Ausreißer enthielten, die Datensätze der erhobenen Stichprobe vollständig und die Voraussetzungen der statistischen Tests erfüllt waren. Alle statistischen Tests und Analysen der Teststärke erfolgten zweiseitig.

Überprüfung der Reliabilität von IntuiBeat-S

Das Vorgehen zur Überprüfung der Reliabilität von IntuiBeat-S war im dritten Experiment identisch mit dem Vorgehen im ersten Experiment, weswegen für genauere Details bezüglich des Vorgehens und der Interpretation an dieser Stelle auf den entsprechenden Absatz des Teilabschnitts 5.1.4.5 des ersten Experiments verwiesen wird.

Überprüfung der Validität von IntuiBeat-S

Das Vorgehen zur Überprüfung der Validität war im dritten Experiment identisch mit dem Vorgehen im zweiten Experiment bezüglich der Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H2 des zweiten Experiments), der Überprüfung der konvergenten und divergenten Validität (Hypothese H3 und H4 des zweiten Experiments), bei der ein Nutzertest mit Rhythmusaufgabe Anwendung fand (d.h. Analysen fanden innerhalb der Faktorstufe *Nutzertest mit Rhythmusaufgabe* des Zwischensubjektfaktors *Art des Nutzertests* statt). Die Auswertung erfolgte dementsprechend durch eine multivariate Varianzanalyse mit Messwiederholung (MANOVA mit Messwiederholung) bzw. mit darauffolgenden univariaten Varianzanalysen mit Messwiederholung (ANOVAs mit Messwiederholung) (siehe Teilabschnitt 5.3.5.3). Für genauere Details wird deswegen

an dieser Stelle auf den entsprechenden Absatz des Teilabschnitts 5.2.3.5 des zweiten Experiments verwiesen. Die Überprüfung der statistischen Voraussetzungen der gerechneten Tests werden im folgenden Ergebnisteil berichtet (siehe Teilabschnitt 5.3.5.1).

5.3.5 Ergebnisse

Im folgenden Abschnitt werden die Ergebnisse bezüglich der in Teilabschnitt 5.3.3 beschriebenen Hypothesen deskriptiv und inferenzstatistisch berichtet. Vor der eigentlichen Datenanalyse wird zunächst auf die Überprüfung der statistischen Voraussetzungen eingegangen. Dabei wurde bei allen abhängigen Variablen und Kontrollvariablen ein metrisches Skalenniveau angenommen.

5.3.5.1 Überprüfung der statistischen Voraussetzungen

Überprüfung von Ausreißern

Zur Überprüfung univariater Ausreißer wurden wie beim ersten Experiment modifizierte z-Werte herangezogen (siehe Teilabschnitt 5.1.5.1 des ersten Experiments). Als univariate Ausreißer wurden in Anlehnung an Iglewicz und Hoaglin (1993) Werte identifiziert, deren absoluter modifizierter z-Wert größer als 3.5 lag. Es hätte auf diese Weise lediglich ein Wert bei AV_SEA ($z = 3.87$) und ein Wert bei AV_PhysischeEffizienz ($z = 3.59$) bei der weniger intuitiv benutzbaren Software ausgeschlossen werden müssen, worauf aufgrund der kleinen Stichprobe ($N = 10$) verzichtet wurde.

Zur Überprüfung multivariater Ausreißer wurde vor Berechnung der multivariaten Varianzanalyse zur Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H1) die Mahalanobis-Distanz herangezogen, da diese ein häufig angewendetes Verfahren zur Identifikation multivariater Ausreißer darstellt (De Maesschalck et al., 2000; Hadi, 1992). Anhand eines Vergleichs der Mahalanobis-Distanz mit den kritischen Werten der Chi-Quadrat-Verteilung (mit $\alpha = .001$ und bei der Gleichsetzung der Freiheitsgrade mit der Anzahl der verwendeten sechs Evaluationsmethoden, die das Ausmaß an intuitiver Benutzung in diesem Experiment operationalisierten), wurde festgestellt, dass sich keine multivariaten Ausreißer in den Daten befanden.

Überprüfung der Voraussetzung der Normalverteilung

Die univariate Normalverteilung der abhängigen Variablen wurde mittels Kolmogorov-Smirnov-Tests ($p \geq .05$, siehe Field, 2017) und Sichtprüfung anhand eines Q-Q-Diagramms geprüft. Dabei konnte im Rahmen der Überprüfung der Unterschiede bei der intuitiven Gestaltung der CUIs (Hypothese H1) und der Überprüfung der konvergenten und divergenten Validität (Hypothesen H2 und H3) bei allen abhängigen Variablen bei beiden Ausprägungen der unabhängigen Variable festgestellt werden, dass diese als normalverteilt einzustufen sind.

Aufgrund der Tatsache, dass mit dem für die statistische Auswertung genutzten, IBM SPSS Statistics 25 kein multivariater Shapiro-Wilk-Test berechnet werden konnte, wurde

eine Überprüfung der multivariaten Normalverteilung der abhängigen Variablen lediglich näherungsweise durch Überprüfung der univariaten Normalverteilung vorgenommen, so wie es eben beschrieben wurde. Obwohl die Normalverteilungsannahme bei den einzelnen Variablen nicht verletzt war, erfolgte die multivariate Varianzanalyse unter Verwendung der Pillai-Spur zur Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H1), da diese nicht nur als robust gegenüber Modellverletzungen gilt, sondern auch eine große statistische Power bei kleinen Stichproben besitzt (Tabachnick et al., 2007).

Überprüfung der Voraussetzung von Linearität und Sicherstellung fehlender Multikollinearität

Die Linearität der abhängigen Variablen wurde für jede Ausprägung der unabhängigen Variable im Rahmen der Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H1) und der Überprüfung der konvergenten und divergenten Validität (Hypothesen H2 und H3) univariat anhand einer Streudiagrammmatrix überprüft. Die, für die multivariaten Varianzanalysen zur Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H1) außerdem erforderliche, fehlende Multikollinearität wurde jeweils mithilfe von Pearson-Produkt-Moment-Korrelationen zwischen den abhängigen Variablen innerhalb der beiden Ausprägungen der unabhängigen Variable geprüft. Aufgrund der Tatsache, dass alle Korrelationen nicht übermäßig hoch lagen ($|r| < .8$), wurde Multikollinearität ausgeschlossen (Berry et al., 1985; Rockwell, 1975).

Überprüfung der Sphärizität

Die Sphärizität zwischen den beiden Ausprägungen der unabhängigen Variable musste nicht mithilfe von Mauchly-Tests im Rahmen der Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H1) berechnet werden, da der Innersubjektfaktor nur zwei Stufen betrug (siehe Field, 2017).

5.3.5.2 Überprüfung der Reliabilität

Wie erwartet zeigte die Analyse der Testhalbierungs-Reliabilität der mittleren Rhythmusabweichungen über alle Versuchspersonen hinweg Guttman Testhalbierungs-Koeffizienten von $r_{Rhinoceros3D} = .997$ (d.h. weniger intuitiv benutzbare Software) und $r_{Fusion360} = .994$ (d.h. stärker intuitiv benutzbare Software), was laut Döring und Bortz (2016) als eine hohe Reliabilität von IntuiBeat-S interpretiert werden kann.

5.3.5.3 Überprüfung der Validität

Tabelle 5.9. *Deskriptive Daten und Teststatistik des dritten Experiments zur Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs (Hypothese H1).*

	Intuitiv ↓ Rhinoceros 3D		Intuitiv ↑ Fusion 360		Teststatistik
	M	*SD*	*M*	*SD*	
AV_IntuiBeat-S [ms]	107.28	31.71	69.59	26.92	$F(1,9) = 11.80$ $p = .007$ $\eta^2 = .57$
AV_CHAI [s]	.17	.13	.37	.13	$F(1,9) = 13.20$ $p = .005$ $\eta^2 = .60$
AV_Effektivität [%]	66.66	9.33	56.49	12.63	$F(1,9) = 3.39$ $p = .099$ $\eta^2 = .27$
AV_QUESI [1-5]	2.67	.53	2.97	.88	$F(1,9) = 1.62$ $p = .235$ $\eta^2 = .15$
AV_NASA [0-100]	55.89	15.55	44.20	12.25	$F(1,9) = 4.83$ $p = .056$ $\eta^2 = .35$
AV_SEA [0-220]	110.39	31.89	76.39	24.14	$F(1,9) = 7.55$ $p = .023$ $\eta^2 = .46$
AV_PhysischeEffizienz [%]	139.50	41.44	136.70	49.59	-
AV_ZeitlicheEffizienz [s]	712.70	95.26	626.00	156.80	-

- : Es erfolgte keine inferenzstatistische Auswertung, da es sich hierbei um divergente Quasi-Außenkriterien handelte, die keine Unterschiede in der intuitiven Gestaltung der CUIs zeigen sollten.

Überprüfung einer möglichen Konfundierung durch die Vorerfahrung bei der Nutzung von CUIs

Da beide Softwarepakete in einem Within-Subjects Design miteinander verglichen wurden, konnte eine mögliche Konfundierung durch die individuelle Vorerfahrung durch die damit verbundene Parallelisierung umgangen werden (siehe Döring & Bortz, 2016). Nichtsdestotrotz kann die erhobene Vorerfahrung bei der Nutzung von CUIs (Wertebereich: 0 - 5) dennoch einen ungewollten Einfluss auf den Vergleich der beiden Ausprägungen der unabhängigen Variable haben. Es wurde dementsprechend mithilfe von Pearson-Produkt-Moment-Korrelationen sichergestellt, dass kein linearer Zusammenhang zwischen der Vorerfahrung und den das Ausmaß intuitiver Benutzung abbildenden Evaluationsmethoden

innerhalb der beiden Ausprägungen der unabhängigen Variable besteht. Die Vorerfahrung bei der Nutzung von CUIs wurde dementsprechend nicht als Kovariate in den folgenden Varianzanalysen berücksichtigt (siehe Döring & Bortz, 2016; Field, 2017).

Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs

Eine multivariate Varianzanalyse zeigte erwartungsgemäß einen signifikanten Effekt der unabhängigen Variable *unterschiedlich intuitiv benutzbare Software* (d.h. weniger intuitiv benutzbare Software vs. stärker intuitiv benutzbare Software) auf das durch die konvergenten Evaluationsmethoden erfasste Ausmaß an intuitiver Benutzung, $F(6,4) = 11.21$, $p = .018$, Pillai-Spur = .944, $\eta_p^2 = .94$. Laut J. Cohen (1988) handelt es sich es hierbei um einen großen Effekt ($\eta_p^2 \geq .14$).

Die weitere univariate varianzanalytische Auswertung der einzelnen, als konvergente Quasi-Außenkriterien fungierenden, Evaluationsmethoden (H1.A bis H1.F) ergab entgegen der Erwartungen keine für die Quasi-Außenkriterien AV_NASA, AV_QUESI und AV_Effektivität signifikanten Effekte. Laut J. Cohen (1988) lagen jedoch bei allen konvergenten Quasi-Außenkriterien große Effekte vor ($\eta^2 \geq .14$). Eine konservative post-hoc Analyse der Teststärke mithilfe von G*Power mit $df = 1$, sowie der Annahme einer Nullkorrelation innerhalb des Innersubjektfaktors ergab bei einer angenommenen hohen ($\eta^2 = .35$ bezüglich AV_NASA) und bei einer weniger hohen Effektstärke ($\eta^2 = .15$ bezüglich AV_QUESI) eine hinreichend große Teststärke (1 - $\beta = .83$) bzw. eine zu geringe Teststärke (1 - $\beta = .40$) im Zuge der Auswertung der abhängigen Variablen AV_NASA, AV_QUESI und AV_Effektivität.

Aufgrund der Tatsache, dass der Unterschied zwischen den unterschiedlich intuitiv benutzbaren CUIs jedoch sowohl multivariat als auch bei den anderen beiden, als konvergente Quasi-Außenkriterien fungierenden, Evaluationsmethoden festgestellt werden konnte, kann die Manipulation (d.h. CUIs unterscheiden sich in ihrer intuitiven Benutzung) als erfolgreich interpretiert werden. Die abhängigen Variablen AV_SEA und AV_CHAI können daher auch empirisch als konvergente Quasi-Außenkriterien (d.h. beide AVs konnten den Unterschied zwischen den unterschiedlich intuitiv benutzbaren CUIs identifizieren) für die Meta-Evaluation von IntuiBeat-S bestätigt werden.

Die anderen drei Evaluationsmethoden AV_NASA, AV_QUESI und AV_Effektivität sollen jedoch dennoch ebenfalls als konvergente Quasi-Außenkriterien für die Überprüfung der konvergenten Validität (Hypothese H2) herangezogen werden, da bei den ermittelten Effektgrößen davon auszugehen ist, dass auch diese Variablen den Unterschied bezüglich des Ausmaßes an intuitiver Benutzung bei einer etwas größeren Stichprobe (d.h. a priori Analyse der Teststärke mithilfe von G*Power liefert eine Stichprobengröße von $N = 13$ bei einem 1 - $\beta = .80$ und einem $\eta^2 = .15$) hätten entdecken können. Wie bereits angesprochen konnten im Rahmen des Projekts jedoch nur 10 Personen für diese Studie ausgewertet werden. Die univariate varianzanalytische Auswertung der beiden Ausprägungen der unabhängigen Variable bezüglich IntuiBeat-S ergab erwartungsgemäß ebenfalls einen signifikanten Effekt (siehe Tabelle 5.1), der laut J. Cohen (1988) ebenfalls als groß ($\eta^2 \geq .14$) interpretiert werden kann. Die deskriptiven Daten der multivariaten und univariaten Varianzanalysen sind Tabelle 5.9 zu entnehmen.

Überprüfung der konvergenten Validität

Wie erwartet, zeigten die Pearson-Produkt-Moment-Korrelationen (siehe Tabelle 5.10) zwischen den durch IntuiBeat-S erhobenen mittleren Rhythmusabweichungen und den Testwerten der anderen, für die Meta-Evaluation fungierenden, konvergenten Quasi-Außenkriterien zur Überprüfung der konvergenten Validität (Hypothese H2) bei beiden Ausprägungen der unabhängigen Variable statistisch signifikante Korrelationskoeffizienten von $|r| > .50$ ($p < .05$), weswegen hier laut J. Cohen (1988) von hohen linearen Zusammenhängen ($|r| \geq .5$) zwischen den Evaluationsmethoden und IntuiBeat-S (H2.A bis H2.E) gesprochen werden kann.

Tabelle 5.10. *Pearson-Produkt-Moment-Korrelationen zwischen summativen Evaluationsmethoden für intuitive Benutzung (d.h. konvergente Quasi-Außenkriterien) und IntuiBeat-S zur Überprüfung der konvergenten Validität (H2) im Zuge des dritten Experiments.*

	AV_IntuiBeat-S: Intuitiv ↓ Rhinoceros 3D	AV_IntuiBeat-S: Intuitiv ↑ Fusion 360
AV_CHAI	$r(10) = -.67$ $p = .036$	$r(10) = -.83$ $p = .003$
AV_Effektivität	$r(10) = -.67$ $p = .034$	$r(10) = -.69$ $p = .028$
AV_QUESI	$r(10) = -.73$ $p = .017$	$r(10) = -.64$ $p = .046$
AV_NASA	$r(10) = .77$ $p = .009$	$r(10) = .70$ $p = .024$
AV_SEA	$r(10) = .73$ $p = .017$	$r(10) = .77$ $p = .010$

Überprüfung der divergenten Validität

Tabelle 5.11. *Pearson-Produkt-Moment-Korrelationen zwischen summativen Evaluationsmethoden, die nicht intuitive Benutzung messen (d.h. divergente Quasi-Außenkriterien) und IntuiBeat-S zur Überprüfung der divergenten Validität (H3) im Zuge des dritten Experiments.*

	AV_IntuiBeat-S: Intuitiv ↓ Rhinoceros 3D	AV_IntuiBeat-S: Intuitiv ↑ Fusion 360
AV_PhysischeEffizienz	$r(10) = -.17$ $p = .650$	$r(10) = .48$ $p = .162$
AV_ZeitlicheEffizienz	$r(10) = -.11$ $p = .754$	$r(10) = .56$ $p = .094$

Wie erwartet, zeigten die Pearson-Produkt-Moment-Korrelationen (siehe Tabelle 5.11) zwischen den durch IntuiBeat-S erhobenen mittleren Rhythmusabweichungen und den

Testwerten der, nicht intuitive Benutzung evaluierenden, divergenten Methoden (d.h. physische Effizienz bei der Systemnutzung und zeitliche Effizienz bei der motorischen Handlungsdurchführung) zur Überprüfung der divergenten Validität (Hypothese H3) bei beiden Ausprägungen der unabhängigen Variable allesamt (H3.A und H3.B) statistisch nicht signifikante lineare Zusammenhänge mit IntuiBeat-S ($p \geq .05$).

5.3.6 Diskussion

Im vorliegenden dritten Experiment wurde die wissenschaftliche Güte von IntuiBeat-S bezüglich des Gütekriteriums *Validität* im Vergleich mit weiteren, als konvergente und divergente Quasi-Außenkriterien fungierenden, summativen Evaluationsmethoden bei der Nutzung einer weniger intuitiv benutzbaren (Rhinoceros 3D, Robert McNeel & Associates, 2017) und einer stärker intuitiv benutzbaren (Fusion 360, Autodesk, 2017a) Software unter Berücksichtigung einer, im Vergleich zu den vorherigen Experimenten (d.h. Studierende der Mensch-Computer-Systeme und der Medienkommunikation der Universität Würzburg), heterogeneren Stichprobe (d.h. Studierende diverser MINT-Fächer aus verschiedenen Institutionen, die erwartungsgemäß über mehr Vorerfahrung bei der Nutzung von CUIs verfügen) und anderer Aufzeichnungshardware (d.h. quelloffenes Fußpedal „Taktschuh“ anstelle eines Fußpedals eines bestimmten Herstellers) empirisch geprüft. Auf diese Weise konnten die Ergebnisse der ersten beiden Experimente auch unter Berücksichtigung einer heterogeneren Stichprobe und unterschiedlicher Hardware bestätigt werden.

Anhand der statistischen Tests wurde zunächst überprüft, inwiefern die Unterschiede in der intuitiven Gestaltung der CUIs mithilfe der erhobenen konvergenten Quasi-Außenkriterien und IntuiBeat-S festgestellt werden können (Hypothese H1). Abschließend wurde untersucht, inwiefern IntuiBeat-S konvergente Validität (Hypothese H2) und divergente Validität (Hypothese H3) unter Berücksichtigung konvergenter und divergenter Quasi-Außenkriterien attestiert werden kann. Im folgenden Verlauf werden nun die Ergebnisse bezüglich der Reliabilität von IntuiBeat-S und die einzelnen Hypothesen bezüglich der Validität unter Berücksichtigung der festgestellten Ergebnisse diskutiert.

5.3.6.1 Überprüfung der Reliabilität

Die, in Form eines Guttman Reliabilitätskoeffizienten berechnete, Testhalbierungsreliabilität von $r_{Rhinoceros3D} = .997$ (d.h. weniger intuitiv benutzbare Software) und $r_{Fusion360} = .994$ (d.h. stärker intuitiv benutzbare Software) lag im Vergleich zum ersten Experiment in einem ähnlichen Wertebereich (d.h. erstes Experiment: $r_{Fusion360} = .983$ und $r_{SketchUp} = .980$), obwohl dort anstelle des im dritten Experiment verwendeten quelloffenen Pedals „Taktschuh“, ein Pedal der Firma Sodial verwendet wurde. Unter Berücksichtigung der Tatsache, dass auch der im Rahmen des dritten Experiments berechnete Reliabilitätskoeffizient laut Döring und Bortz (2016) als hoch zu interpretieren ist und sich auch in einem ähnlichen Wertebereich wie die Ursprungsstudie zur Rhythmusmethode ($r_{erstesExperiment} = .96$ und $r_{zweitesExperiment} = .72$, siehe Park & Brünken, 2015) und die Folgestudie ($r = .96$, siehe Korbach et al., 2018) befindet, kann IntuiBeat-S als summative Evaluationsmethode für intuitive Benutzung auch bei Verwendung des quelloffenen Pedals „Taktschuh“

Reliabilität attestiert und damit das entsprechende Gütekriterium als gegeben angesehen werden.

5.3.6.2 Überprüfung der Validität

Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs

Wie erwartet konnten die, als konvergente Quasi-Außenkriterien fungierenden, Evaluationsmethoden SEA und CHAI signifikant ein höheres Ausmaß an intuitiver Benutzung bezüglich der stärker intuitiv benutzbaren Software im Vergleich zur weniger intuitiv benutzbaren Software feststellen. Es ließen sich hierbei durchweg hohe Effekte beobachten, wodurch sich diese Evaluationsmethoden auch empirisch als konvergente Quasi-Außenkriterien für die Meta-Evaluation von IntuiBeat-S qualifizieren und die in diesem Zusammenhang aufgestellten Hypothesen (H1.B und H1.F) als bestätigt angesehen werden können. Entgegen der Erwartungen konnte jedoch durch die Messung der Effektivität, die Erhebung des QUESI und des NASA-RTLX kein signifikantes Ergebnis bezüglich dieses gerichteten Unterschiedes ermittelt und die damit verbundenen Hypothesen H1.C, H1.D und H1.E somit nicht bestätigt werden. Aufgrund der Tatsache, dass auch in diesen Fällen die Effektstärke hoch war und man mit einer etwas größeren Stichprobe auch mit hoher Wahrscheinlichkeit bezüglich dieser Maße einen signifikanten Unterschied zwischen den beiden Ausprägungen der unabhängigen Variablen hätte feststellen können, wurden auch diese Maße als konvergente Quasi-Außenkriterien für die Überprüfung der konvergenten Validität verwendet. Wie erwartet, konnte mithilfe von IntuiBeat-S signifikant ein höheres Ausmaß an intuitiver Benutzung bezüglich der stärker intuitiv benutzbare Software im Vergleich zur weniger intuitiv benutzbaren Software feststellt und die in diesem Zusammenhang aufgestellte Hypothese (H1.A) bestätigt werden (siehe Teilabschnitt 5.12).

Überprüfung der konvergenten Validität

Wie erwartet zeigten sich hohe signifikante lineare Zusammenhänge (d.h. konvergente Validitätskoeffizienten) zwischen IntuiBeat-S und allen, als konvergente Quasi-Außenkriterien fungierenden, Evaluationsmethoden, weswegen alle in diesem Zusammenhang aufgestellten Hypothesen bestätigt werden können (H2.A bis H2.E, siehe Teilabschnitt 5.12).

Überprüfung der divergenten Validität

Wie erwartet zeigten sich nicht signifikante lineare Zusammenhänge (d.h. divergente Validitätskoeffizienten) zwischen IntuiBeat-S und den, nicht intuitive Benutzung evaluierenden, divergenten Methoden, weswegen alle in diesem Zusammenhang aufgestellten Hypothesen bestätigt werden können (H3.A und H3.B, siehe Teilabschnitt 5.12). Dennoch besteht in der Operationalisierung der physischen Effizienz als Anzahl der Klicks bei der Systemnutzung die Gefahr, dass dieses Maß mit Evaluationsmethoden, die die intuitive Benutzung abbilden, bei einer stärker intuitiv benutzbaren Software ungewollt korrelieren könnte. Diese Gefahr konnte mithilfe von Pearson-Produkt-Moment-Korrelationen ($p > .05$) jedoch ausgeschlossen werden.

5.3.7 Schlussfolgerung

Zusammenfassend kann festgehalten werden, dass die wissenschaftliche Güte von IntuiBeat-S als summative Evaluationsmethode für intuitive Benutzung hinsichtlich der Gütekriterien *Reliabilität* und *Validität* auch unter Berücksichtigung einer, im Vergleich zu den vorherigen Studien, heterogeneren Stichprobe und eines anderen USB-Pedals empirisch bestätigt werden konnte (siehe Teilabschnitt 5.12). Aufgrund des Messwiederholungsdesigns können darüber hinaus Konfundierungen durch Unterschiede in der Vorerfahrung bei der Nutzung von CUIs mit hoher Wahrscheinlichkeit ausgeschlossen werden. Die Effekt- und Teststärken lagen dabei überwiegend im oberen Bereich. Jedoch konnten bezüglich der Fragebögen NASA-RTLX und QUESI, sowie des objektiven Leistungsmaßes der Effektivität keine signifikanten Unterschiede zwischen den beiden Ausprägungen der unabhängigen Variable festgestellt werden. Dies lässt sich zum einen offensichtlich auf den geringen Stichprobenumfang des Experiments ($N = 10$) zurückführen, was aber aufgrund der auch bei diesen Maßen beobachteten hohen Effektgrößen weniger kritisch ist, da durch eine marginal größere Stichprobe ($N = 13$ aufgrund einer a priori Analyse der Teststärke mithilfe von G*Power bei einem 1 - $\beta = .80$ und einem $\eta^2 = .15$) mit hoher Wahrscheinlichkeit signifikante Unterschiede hätten ermittelt werden können. Zum anderen liefert dieses Ergebnis noch eine weitere, viel wichtigere Erkenntnis.

Tabelle 5.12. *Übersicht der mithilfe des dritten Experiments bestätigten Hypothesen im Zuge der Meta-Evaluation von IntuiBeat-S.*

Hypothese	Experiment 3
(H1) Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs:	
- (A) IntuiBeat-S	✓
- (B) SEA	✓
- (C) NASA-RTLX	✗
- (D) QUESI	✗
- (E) Effektivität	✗
- (F) CHAI	✓
(H2) Überprüfung der konvergenten *Validität:*	
- (A) SEA	✓
- (B) NASA-RTLX	✓
- (C) QUESI	✓
- (D) Effektivität	✓
- (E) CHAI	✓
(H3) Überprüfung der divergenten Validität:	
- (A) Physische Effizienz	✓
- (B) Zeitliche Effizienz	✓

Da sowohl die beiden objektiven Evaluationsmethoden CHAI und IntuiBeat-S einen Unterschied bezüglich des Ausmaßes an intuitiver Benutzung entdecken konnten, es aber bei einigen subjektiven Methoden (d.h. QUESI und NASA-RTLX) und den anderen objektiven Methoden (d.h. Effektivität) an der nötigen Sensitivität mangelte, lässt sich daraus eine

höhere Sensitivität der objektiven Maße IntuiBeat-S und CHAI folgern. Diese Erkenntnis unterstützt den aktuellen Benchmarkstatus der CHAI-Methode und unterstreicht das Potential von IntuiBeat-S als künftiger Benchmark im Bereich der summativen Evaluation intuitiver Benutzung weiter. Ein möglicher Grund für die geringere Sensitivität des objektiven Hauptaufgabenleistungsmaßes der Effektivität kann die, in Teilabschnitt 3.6.2.3 angesprochene, Anwendung von Adaptionsstrategien des Nutzers sein (siehe F. Chen et al., 2016), die er entweder bei geringer mentaler Beanspruchung (d.h. Effektivität spiegelt bei geringer mentaler Beanspruchung diese nicht notwendigerweise korrekt wieder, da der Nutzer aufgrund seiner freien Ressourcen diese durch erhöhte Anstrengung kompensieren kann, die tatsächliche Effektivität dadurch nach oben verzerrt wird und keine Unterschiede in der Effektivität aufgrund eines Deckeneffekts mehr feststellbar sind) oder hoher mentaler Beanspruchung (d.h. Effektivität spiegelt bei hoher mentaler Beanspruchung diese nicht notwendigerweise korrekt wieder, da der Nutzer, um seine wenigen vorhandenen Ressourcen zu sparen, seine Effektivität durch Unterlassen bestimmter Aktionen nach unten verzerrt und keine Unterschiede aufgrund eines Bodeneffekts in der Effektivität mehr feststellbar sind) zu seinen Gunsten anwendet (Eggemeier et al., 1991; O'Donnell & Eggemeier, 1986; Wierwille & Eggemeier, 1993). Laut S. G. Hart und Wickens (1990) liefert dies eine Erklärung, warum die wissenschaftliche Güte von Hauptaufgabenleistungsmaßen bei der Erfassung von mentaler Beanspruchung im mittleren Bereich am höchsten ist.

Ein Grund für die mögliche geringere Sensitivität von subjektiven Maßen können die, in Teilabschnitt 3.6.1.6 angesprochenen, allgemeinen Limitationen von subjektiven Maßen sein, wie beispielsweise die Anfälligkeit gegenüber einem Primacy-Recency-Effekt durch die retrospektive Erfassung oder anderen bekannten Phänomenen (z.B. Verzerrung von subjektiven Maßen durch Einstellungen des Nutzers, siehe Cockburn et al., 2015). Es lässt sich dementsprechend vermuten, dass sich bei Maßen wie dem QUESI (siehe Teilabschnitt 3.6.1) und dem NASA-RTLX, der das Gefühl von Flüssigkeit anhand der mentalen Arbeitsbelastung aus einer Belastungs- und einer Beanspruchungsperspektive betrachtet (siehe Teilabschnitt 3.6.1), sich aufgrund der differenzierteren Erfassungsweise Verzerrungen stärker niederschlagen können. Dieser Aspekt könnte möglicherweise erklären, warum es zu nicht signifikanten Ergebnisse bezüglich beider subjektiver Maße gekommen ist. An dieser Stelle soll jedoch angemerkt werden, dass diese Alternativerklärungen aufgrund des geringen Stichprobenumfangs mit Vorsicht zu genießen und auch bei den beiden Vorgängerexperimenten anwendbar sind.

5.4 Allgemeine Diskussion und Zusammenfassung

Das Ziel des vorliegenden Kapitels war es zu demonstrieren, dass die Evaluationsmethode *IntuiBeat-S*, die eine Adaption der von Park und Brünken (2015) entwickelten Rhythmusmethode für den HCI-Bereich darstellt, auch in diesem Anwendungsfeld wissenschaftlich valide summative Evaluationsergebnisse auf zeitlich effiziente Weise bereitstellen kann. Laut des in Abschnitt 2.2 vorgestellten Drei-Instanzen Modells des Geistes kann zwischen intuitiver und reflektierender kognitiver Verarbeitung anhand des Bedarfs kognitiven Abkoppelns und des damit einhergehenden Gefühls von Flüssigkeit differenziert werden. Das Ausmaß an intuitiver Benutzung lässt sich demnach objektiv anhand der mentaler Beanspruchung und subjektiv anhand des metakognitiven Gefühls von Flüssigkeit feststellen.

Tabelle 5.13. *Übersicht der, mithilfe der Experimente 1 bis 3, bestätigten Hypothesen im Zuge der Meta-Evaluation von IntuiBeat-S.*

Hypothese	Experiment 1	Experiment 2	Experiment 3
Überprüfung der Intrusion durch die Art des Nutzertests:			
- Haupteffekt des Innersubjekt-faktors *unterschiedlich intuitiv benutzbare Software*	-	✓	-
- Kein Haupteffekt des Zwischen-subjektfaktors *Art des Nutzertests*	-	✓	-
- Kein Interaktionseffekt des Zwischensubjektfaktors und des Innersubjekfaktors	-	✓	-
Überprüfung der Unterschiede in der intuitiven Gestaltung der CUIs:			
- IntuiBeat-S	✓	✓	✓
- SEA	✓	✓	✓
- NASA-RTLX	✓	✓	✗
- QUESI	✓	✓	✗
- Effektivität	✗	✓	✗
- CHAI	✓	✓	✓
Überprüfung der konvergenten *Validität:*			
- SEA	✓	✓	✓
- NASA-RTLX	✓	✓	✓
- QUESI	✓	✓	✓
- Effektivität	✓	✓	✓
- CHAI	✓	✓	✓
	✓	✓	✓
Überprüfung der divergenten Validität:			
- Physische Effizienz	✓	✓	✓
- Zeitliche Effizienz	✓	✓	✓
Überprüfung der zeitlichen Anwendungseffizienz	-	✓	-

Da IntuiBeat-S eine Inhibitionsaufgabe als Rhythmuszweitaufgabe zur Erfassung mentaler Beanspruchung nutzt und Inhibitionsprozesse im Allgemeinen als universelle Indikatoren mentaler Beanspruchung gelten (J. D. Cohen et al., 1997; A. Miyake et al., 2000; Park & Brünken, 2015), kann diese Rhythmuszweitaufgabe laut dem Modell multipler Ressourcen theoretisch eine hohe Interferenz erzeugen. Schwankungen in der, durch die Systeminteraktion (d.h. Hauptaufgabe) verursachten, mentalen Beanspruchung können auf diese Weise sichtbar gemacht werden, ohne dabei gleichzeitig intrusiv zu sein. Dies konnte bereits mehrfach empirisch durch die Vorgängerstudien zur Rhythmusmethode (siehe Korbach et al., 2018; Park & Brünken, 2015) nachgewiesen werden. Dementsprechend sollte IntuiBeat-S das Ausmaß an intuitiver Benutzung objektiv anhand der mentalen Beanspruchung erfassen können und dabei mit weiteren konvergenten Evaluationsmethoden auf Basis mentaler

Beanspruchung und dem damit verbundenen Gefühl von Flüssigkeit in Zusammenhang stehen. Divergente Evaluationsmethoden (z.B. motorische Effizienz), die sich nicht der objektiven Erfassung mentaler Beanspruchung oder der subjektiven Erfassung des metakognitiven Gefühls von Flüssigkeit widmen, sollten damit nicht in Verbindung stehen.

Die Ergebnisse der, in diesem Kapitel beschriebenen, drei Experimente konnten erwartungsgemäß bestätigen (siehe Tabelle 5.13), dass IntuiBeat-S wissenschaftliche und praktische Güte (d.h. in Form zeitlicher Anwendungseffizienz) als objektive Methode zur Evaluation von User Interfaces (speziell 3D-CUIs) bezüglich des Ausmaßes an intuitiver Benutzung attestiert werden kann. IntuiBeat-S kann daher zwischen unterschiedlich intuitiv benutzbaren User Interfaces (speziell 3D-CUIs) differenzieren. Beide in diesem Zusammenhang aufgestellten Forschungsfragen (Forschungsfrage 1 und summativer Aspekt von Forschungsfrage 3) konnten demzufolge beantwortet werden (siehe Abschnitt 4.3). Im Folgenden werden die Befunde bezüglich des Nachweises wissenschaftlicher Güte und der zeitlichen Anwendungseffizienz zusammengefasst betrachtet und kritisch diskutiert, einhergehend mit Bezugnahme zur Forschungsliteratur.

Aufgrund der Tatsache, dass IntuiBeat-S ein hohes Maß an Automatisierung bietet, wurde das wissenschaftliche Gütekriterium der Objektivität nicht empirisch untersucht und die Erfüllung dieses Kriteriums stattdessen auf theoretische Weise begründet. Da die Durchführung in allen drei Experimenten von verschiedenen Evaluatoren mit unterschiedlicher Vorerfahrung übernommen wurde (d.h. unterschiedliche Studienabschlüsse und damit mit hoher Wahrscheinlichkeit auch unterschiedliche Vorerfahrung) und die Datenerhebung auch innerhalb eines Experiments teilweise durch verschiedene Evaluatoren erfolgte (z.B. im ersten Experiment), kann die Objektivität der Methode auch ohne empirischen Nachweis mit hoher Wahrscheinlichkeit angenommen werden. Dieses Vorgehen wurde bereits in der Einleitung dieses Kapitels begründet. Als Limitation ist jedoch an dieser Stelle zu erwähnen, dass es sich bei den Evaluatoren vorwiegend um junge Evaluatoren handelte, die gerade ihren Bachelor oder Master abgeschlossen hatten oder am Ende ihres Studiums waren. Betrachtet man diesen Aspekt unter Berücksichtigung der signifikanten Ergebnisse der beschriebenen Experimente, muss dies aber keine Einschränkung im engeren Sinne darstellen. Da mithilfe von IntuiBeat-S und der IntuiBeat-Software mit unterschiedlicher Hardware (d.h. USB-Fußpedal eines bestimmten Herstellers oder des USB-Fußpedals „Taktschuh“) sogar verschiedene Evaluatoren mit geringem Kenntnisstand bei der Durchführung einer summativen Evaluation auf Basis eines Nutzertests (d.h. Anzahl der durchgeführten Nutzertests bei den Evaluatoren erwartungsgemäß noch im einstelligen Bereich) erfolgreich wissenschaftlich tragfähige Ergebnisse produzieren konnten. Für zukünftige Forschung wäre es sicherlich spannend zu sehen, ob sich Evaluatoren mit mehr Vorerfahrung mit summativen Nutzertests auch pedantisch an die stark standardisierte Vorgehensweise von IntuiBeat-S halten oder auf Basis ihrer Erfahrung daraus eher ausbrechen möchten und welche Auswirkungen dies auf die Güte (d.h. wissenschaftliche und praktische Güte, bei letzterer auch andere Aspekte neben zeitlicher Anwendungseffizienz) von IntuiBeat-S hat.

Das wissenschaftliche Gütekriterium der Reliabilität von IntuiBeat-S konnte sowohl im ersten Experiment bei der Nutzung eines handelsüblichen USB-Fußpedals, bei dem der Rhythmus vom Nutzer per Fußballen eingegeben wurde, als auch bei der Nutzung des quelloffenen, am Lehrstuhl für Psychologische Ergonomie konstruierten USB-Fußpedals

„Taktschuh“ in Kombination mit der IntuiBeat-Software anhand von hohen Reliabilitätskoeffizienten nachgewiesen werden. Da sich die beiden Reliabilitätskoeffizienten (Experiment 1 und Experiment 3) zudem in einem ähnlichen Wertebereich ($r \geq .9$) befanden (d.h. in allen Versuchsbedingungen), lässt sich ferner schließen, dass für die Anwendung von IntuiBeat-S sowohl handelsübliche USB-Fußpedale als auch das im Rahmen des Projekts 3D-GUIde entwickelte USB-Fußpedal „Taktschuh“ bei der Evaluation von weniger und stärker intuitiv benutzbarer Software verwendet werden können. Diese Tatsache ist praktisch durchaus relevant, da Evaluatoren somit bei der Auswahl ihrer Hardware nicht auf ein bestimmtes Fabrikat beschränkt sind. Auf diese Weise konnte die versprochene Plug und Play Lösung bereitgestellt werden, die nicht, wie die in den Ursprungsstudien zur Rhythmusmethode (Korbach et al., 2018; Park & Brünken, 2015), auf bestimmte Komponenten (z.B. Serial Response Box und E-Prime kompatibles Fußpedal) zurückgreift, die entweder schwer verfügbar oder teuer sind.

Darüber hinaus konnte man durch die Durchführung von IntuiBeat-S mithilfe eines klassischen USB-Fußpedals (Experiment 1), bei dem der Rhythmus per Fußballen eingegeben wurde, und des Pedals „Taktschuh“ (Experiment 3), bei dem die Eingabe des Rhythmus per Ferse erfolgte, zusätzliche Flexibilität bei der Anwendung von IntuiBeat-S demonstrieren. Der Evaluator kann sich dementsprechend auf physische Einschränkungen (z.B. unterschiedliche Ausprägung der Muskulatur, Berücksichtigung von Bewegungseinschränkungen) einstellen und die Erfassung des Rhythmus für den Probanden so angenehm wie möglich gestalten. Auf Grundlage der Ergebnisse sollte es bei künftigen Erhebungen mit hoher Wahrscheinlichkeit auch kein Problem darstellen, zwischen verschiedenen Eingabearten (d.h. Ferse oder Fußballen) zu wechseln, um Ermüdungserscheinungen zu vermeiden. Nichtsdestotrotz ist an dieser Stelle anzumerken, dass die Aufgaben in den Experimenten bewusst kurz (d.h. wenige Minuten) gehalten wurden, um die Ermüdungswahrscheinlichkeit damit zu minimieren. Da es in der Praxis auch vorkommen kann, dass man die Systemnutzung länger evaluieren möchte, sollte die Reliabilität von IntuiBeat-S künftig auch in diesen Situationen untersucht und der Zeitpunkt festgestellt werden, wann und ob die Reliabilität der Messung ermüdungsbedingt abnimmt.

Das wissenschaftliche Gütekriterium der Validität wurde in den beschriebenen drei Experimenten jeweils durch den Vergleich von zwei unterschiedlich intuitiv benutzbaren CUIs (d.h. eines weniger und eines stärker intuitiv) durch Überprüfung der konvergenten und divergenten Validität erfolgreich nachgewiesen (siehe Tabelle 5.13). Zuvor wurde bei den konvergenten, als Quasi-Außenkriterien fungierenden, Evaluationsmethoden und IntuiBeat-S überprüft, ob diese Methoden den Unterschied zwischen den beiden unterschiedlich intuitiv benutzbaren CUIs feststellen und sich diese daher auch empirisch als summative Evaluationsmethoden für intuitive Benutzung qualifizieren können. Die Ergebnisse der drei Experimente zeigten, dass die konvergenten Evaluationsmethoden *Messung der Effektivität* (Experiment 1 und 3) und die beiden Fragebögen *NASA-RTLX* und *QUESI* (Experiment 3) sich nicht durchgängig empirisch als konvergente Quasi-Außenkriterien qualifizieren konnten. Dies kann unter anderem generell auf die etwas zu kleinen Stichprobengrößen in den drei Experimenten zurückgeführt werden. Diese mangelnde Sensitivität ließ sich jedoch unter Berücksichtigung der kleinen Stichproben nicht auf eine Konfundierung durch unterschiedliche Vorerfahrung, unterschiedliche rhythmische Wahrnehmung oder Intrusion aufgrund des als Datengrundlage genutzten Nutzertests mit Rhythmuszweitaufgabe zurückführen.

Wie bereits in der Einleitung dieses Kapitels beschrieben, musste aufgrund des straffen Zeitplans im Projekt 3D-GUIde die Meta-Evaluation von IntuiBeat-S parallel zur qualitativen Datenerhebung in kurzer Zeit erfolgen, weswegen insgesamt nur wenige Versuchspersonen im Rahmen dieser Meta-Evaluation getestet werden konnten. Bezieht man sich jedoch auf Nielsen (2006) liegen die Stichprobengrößen der für die Meta-Evaluation von IntuiBeat-S genutzten Experimente (d.h. $N_{Experiment\_1} = 36$, $N_{Experiment\_2} = 38$, $N_{Experiment\_3} = 10$) in einem praktisch relevanten Umfang. Laut Nielsen (2006) sollte man in der Praxis für die summative Evaluation immer mindestens 20 Nutzer (also viermal so viel Nutzer wie bei formativen Evaluationen, siehe Nielsen & Landauer, 1993) testen, wobei diese Zahl bei Usability-Agenturen häufig aufgrund von Zeitgründen unterschritten wird und nur lediglich 11 Nutzer im Schnitt getestet werden (Nielsen, 2012). Obwohl die kleinen Stichprobengrößen wahrscheinlich dazu führten, dass mithilfe der Messung der Effektivität (Experiment 1 und 3), des QUESI (Experiment 3) und des NASA-RTLX (Experiment 3) keine signifikanten Unterschiede zwischen den zwei unterschiedlich intuitiv benutzbaren CUIs feststellbar waren, konnte dieser Unterschied durchgängig objektiv mit IntuiBeat-S und der CHAI-Methode, sowie subjektiv mit der SEA-Skala festgestellt werden. Auf diese Weise konnte gezeigt werden (alle Experimente), dass diese Methoden erfolgreich zur summativen Evaluation intuitiver Benutzung auch unter Berücksichtigung der in der Praxis üblicherweise genutzten Anzahl an Versuchspersonen eingesetzt werden können und sich damit auch empirisch als konvergente Quasi-Außenkriterien qualifizieren.

Darüber hinaus konnte IntuiBeat-S aufgrund der Verbindung mit Inhibitionsprozessen und der dadurch möglichen modalitätsunabhängigen Erfassung des Ausmaßes an intuitiver Benutzung anhand der mentalen Beanspruchung konvergente und divergente Validität attestiert werden. Die konvergenten Korrelationskoeffizienten lagen bei allen drei Experimenten durchgehend im hohen Bereich ($|r| \geq .5$, siehe J. Cohen, 1988) auch bezüglich der Effektivität, NASA-RTLX und QUESI, die sich nicht durchgängig als valide summative Evaluationsmethoden für intuitive Benutzung qualifizieren konnten. Des Weiteren blieben die Zusammenhänge zwischen IntuiBeat-S und den divergenten Evaluationsmethoden über alle drei Experimente hinweg erwartungsgemäß nicht signifikant.

Zusammenfassend lässt sich daraus folgern, dass auch empirisch bestätigt werden konnte, dass die mit IntuiBeat-S erfasste mentale Beanspruchung allen, in Kapitel 2 aufgeführten, Merkmalen intuitiver Benutzung als zentrales objektives Merkmal zugrunde liegt. Die am Ende des Kapitels formulierte Messdefinition, welche intuitive Benutzung als das Ausmaß beschreibt, mit dem ein Produkt effektiv und mental effizient genutzt werden kann, was mit einem starken metakognitiven subjektiven Gefühl von Flüssigkeit einhergeht, ist dementsprechend im Gegensatz zu den bestehenden Definitionen der einzelnen Forschergruppen in der Lage, über das Konzept der mentalen Beanspruchung auch den metakognitiven Aspekt intuitiver Benutzung zu berücksichtigen, so wie dieser von anderen Wissenschaftler innerhalb der HCI-Forschung zu intuitiver Benutzung (z.B. O'Brien et al., 2010) implizit angesprochen und außerhalb dieser Forschung (z.B. Ackerman & Thompson, 2017; V. Thompson et al., 2011; Topolinski, 2011) sogar differenziert diskutiert wird. Sie ist demzufolge recht gut als Definition geeignet, wenn es um die Evaluation intuitiver Benutzung geht, da sie auf die Nennung gestaltungsrelevanter Aspekte, wie die unbewusste Anwendung von Vorwissen (siehe Teilabschnitt 2.1.1), verzichtet.

Neben der wissenschaftlichen Güte konnte die Methode IntuiBeat-S auch ein, für das Projekt 3D-GUIde, wichtiger Teilaspekt praktischer Güte aufgrund ihrer höheren zeitlichen Anwendungseffizienz gegenüber dem aktuellen objektiven Benchmark im Bereich der summativen Evaluation intuitiver Benutzung, der CHAI-Methode, attestiert werden. Da IntuiBeat-S somit nicht nur durchgängig wissenschaftliche Güte zugestanden werden kann, sondern dieser auch dem aktuellen objektiven Benchmark CHAI bezüglich der zeitlichen Anwendungseffizienz vorzuziehen ist (d.h. benötigt nur ein Fünftel der Anwendungszeit) und als objektive Methode weniger anfällig als subjektive Verfahren (z.B. SEA, NASA-RTLX und QUESI) gegenüber kognitiven Verzerrungen durch die retrospektive Erfassung des Ausmaßes an intuitiver Benutzung ist (siehe Blackler, 2018), konnte IntuiBeat-S problemlos im Projekt 3D-GUIde zur schnellen Evaluation von 3D-Interaktionspattern eingesetzt werden (siehe Burmester et al., 2018).

Nichtsdestotrotz wurden im Rahmen der drei Experimente nur die Domäne der CUIs mit einer beschränkten Auswahl an Aufgaben betrachtet, sowie die CUIs lediglich von Experten auf analytische Weise als weniger oder stärker intuitiv benutzbar eingestuft und der Nachweis der Güte von IntuiBeat-S dadurch auf einen bestimmten Kontext, bestimmte Aufgaben und bestimmte Softwareanwendungen limitiert. Die Untersuchung der externen Validität und damit die Übertragung auf andere Bereiche steht dementsprechend noch aus. Zukünftige Forschung besitzt dadurch die Möglichkeit eine Meta-Evaluation von IntuiBeat-S auch in anderen Bereichen vorzunehmen und damit sowohl die vorhandene wissenschaftliche als auch die zeitliche Anwendungseffizienz von IntuiBeat-S als summative Evaluationsmethode für intuitive Benutzung weiter zu untermauern. Da man sich in dieser Arbeit wegen der Anforderungen des Projekts 3D-GUIde auf die zeitlich ökonomischen Aspekte fokussiert hat, wäre es für zukünftige Forschung sicherlich spannend weitere Nebengütekriterien wie Normierung in die Meta-Evaluation von IntuiBeat-S miteinzubeziehen. Durch die Normierung von IntuiBeat-S könnte Anwendern in der Wissenschaft und der Praxis Orientierungswerte geboten werden, auf deren Grundlage sie das beobachtete Ausmaß an intuitiver Benutzung beim getesteten System interpretieren und einordnen können. Derartige Benchmarks sind meines Wissens nur für subjektive Maße wie beispielsweise den QUESI bekannt (siehe Naumann & Hurtienne, 2010) und fehlen für objektive Maße im Forschungsfeld der intuitiven Benutzung (vergleiche Blackler, 2018). Zusätzlich konnten im Zuge der drei Experimente nur Versuchspersonen getestet werden, die den Rhythmus mit ihrem rechten Fuß eingaben. Zukünftige Experimente sollten auch die wissenschaftliche Güte von IntuiBeat-S bei der Eingabe mit dem linken Fuß untersuchen und überprüfen, ob die Güte allgemein auch durch die Möglichkeit den Fuß während der Systemnutzung zu wechseln, beeinflusst wird.

Aufgrund der Tatsache, dass IntuiBeat-S durch ihr stationäres Setup vom Nutzer verlangt, sitzend den Rhythmus zu klopfen und man dadurch auf die Evaluation von Desktopsystemen und mobilen Geräte im Sitzen beschränkt ist, wäre es nötig zu klären, inwiefern eine Evaluation im Stehen oder Gehen möglich ist. Da eine kontinuierliche Rhythmuseingabe beim Gehen mit hoher Wahrscheinlichkeit alleine schon aufgrund der damit verbundenen Bewegung schwierig ist, sollten zu diesem Zweck auch andere wenig invasive Eingabemöglichkeiten als Möglichkeit zur Rhythmuseingabe erprobt werden. Beispielsweise würden sich hier verschiedene Formen des Finger- oder Handklopfens (z.B. M. J. Albers, 2011; Y. Miyake et al., 2004), Rhythmuseingabe mithilfe diverser Sensoren im Mundbereich wie Druckmessung beim Ausatmen, Messung der Veränderung der Gesichtsmuskulatur, oder

Erfassung der Zunge mithilfe eines Magnetsensors (z.B. Cheng et al., 2014; Huo, Wang, & Ghovanloo, 2008; Kapur, Kapur, & Maes, 2018; Scavone, 2003) eignen, da diese auch großes Potential bei der Evaluation von virtuellen Umgebungen im Mixed Reality Umfeld besitzen. Intuitive Benutzung ist in diesem Bereich noch wenig erforscht (siehe Blackler, 2018) und ihr sollte deswegen zukünftig eine stärkere Beachtung geschenkt werden.

Darüber hinaus hatten die drei Experimente nicht nur die Meta-Evaluation von IntuiBeat-S als summative Evaluationsmethode intuitiver Benutzung als Ziel, sondern auch die qualitative Ableitung von Nutzungsproblemen für die Interaktionspatternentwicklung. Aufgrund des engen Zeitplans im Anwenderprojekt (d.h. es musste innerhalb des ersten Jahres eine summative Evaluationsmethode für intuitive Benutzung entwickelt und validiert werden, die im Vergleich zu bekannten Methoden eine hohe zeitliche Anwendungseffizienz aufweisen konnte) und der Wichtigkeit dieser qualitativen Informationen für den weiteren Projektablauf, wurden bei den drei Experimenten nicht dieselben Stimuli (d.h. CUIs) und Evaluatoren verwendet, da kontinuierlich neue qualitative Informationen von diversen CUIs für die Entwicklung von 3D-CUI-Interaktionspatterns benötigt wurden. Die Meta-Evaluation von IntuiBeat-S musste daher parallel laufen. Da die Experimente daher in ihrem Kern eher qualitativ ausgerichtet waren und die Meta-Evaluation von IntuiBeat-S nur begleitend erfolgte, weißen die drei Experimente geringere Stichprobenumfänge und eine weniger standardisiertes Vorgehen auf, als es bei einer Meta-Evaluation einer summativen Evaluationsmethode wünschenswert gewesen wäre. Künftige Meta-Evaluationen von IntuiBeat-S sollten daher versuchen die Experimente unter Verwendung der gleichen Versuchsbedingungen (d.h. gleiche CUIs, Aufgaben und Stichprobe) direkt zu replizieren anstelle des im Rahmen dieser Arbeit verfolgten Ansatzes der konzeptuellen Replikation. Der benötigte Stichprobenumfang kann hierfür mithilfe einer Poweranalyse auf Basis der im Rahmen der drei Experimente ermittelten Effektstärken a priori abgeschätzt werden.

Abschließend lässt sich zusammenfassen, dass die eben beschriebenen drei Experimente sich nur der Meta-Evaluation von IntuiBeat-S widmeten, um dadurch die Möglichkeit zu schaffen, wissenschaftlich tragfähige, sowie zeitlich effiziente Ergebnisse bei der summativen Evaluation intuitiver Benutzung zu erzielen. Da eine weitere Anforderung des Projekts 3D-GUIde darin bestand, auch eine formativ zeitlich effizient einsetzbare Evaluationsmethode für intuitive Benutzung bereitzustellen, und man laut aktueller Forschungsliteratur (siehe Blackler, 2018) aufgrund des Mangels an dedizierten Methoden lediglich auf einen gewöhnlichen Nutzertest mit Think-Aloud.Protokoll zurückgreifen muss, soll sich das folgende Kapitel der Meta-Evaluation von IntuiBeat-F und damit der formativen Evaluation intuitiver Benutzung widmen.

6 Güte von IntuiBeat-F für die formative Evaluation intuitiver Benutzung

Das vorherige Kapitel demonstrierte die wissenschaftliche Güte und zeitliche Anwendungseffizienz von IntuiBeat-S. Mithilfe dieses Kapitels soll nun die wissenschaftliche Güte und zeitliche Anwendungseffizienz von IntuiBeat-F durch die Beantwortung der zweiten Forschungsfrage und des formativen Aspekts der dritten Forschungsfrage (d.h. betrifft IntuiBeat-F) betrachtet werden:

Forschungsfrage 2 Wie hoch ist die wissenschaftliche Güte von IntuiBeat-F als formative Evaluationsmethode für intuitive Benutzung, beurteilt anhand der formalen Hauptgütekriterien Gründlichkeit, Gültigkeit und Zuverlässigkeit?

Forschungsfrage 3 Wie hoch ist die zeitliche Anwendungseffizienz von IntuiBeat-S und IntuiBeat-F im Vergleich zu bereits vorhandenen Evaluationsmethoden für intuitive Benutzung?

Zur Beantwortung der zweiten Forschungsfrage und des formativen Aspekts der dritten Forschungsfrage wurden in vier Experimenten unterschiedlich intuitiv benutzbare Softwareanwendungen einzeln formativ evaluiert. Hierbei repräsentierten *IntuiBeat-F* (siehe Abschnitt 3.2.3) und ein als Quasi-Außenkriterium fungierender *Nutzertest mit retrospektivem Think-Aloud-Protokoll* (siehe Abschnitt 3.5) die Ausprägungen der unabhängigen Variable *Art der formativen Evaluationsmethode*, welche bezüglich der abhängigen Variablen *Gründlichkeit* (d.h. Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der vorhandenen realen Nutzungsprobleme), *Gültigkeit* (d.h. Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der von der jeweiligen Evaluationsmethode gefundenen Nutzungsprobleme) und *zeitliche Anwendungseffizienz* miteinander verglichen wurden (siehe Teilabschnitt 3.3).

Theoretische Begründung zur Vernachlässigung der Zuverlässigkeit

Aufgrund der Tatsache, dass Hartson et al. (2001) empfehlen, sich der Zuverlässigkeit einer formativen Evaluationsmethode erst dann zu widmen, nachdem sowohl die Gründlichkeit als auch die Gültigkeit der Methode bestätigt werden konnte, wurde im Rahmen dieser Arbeit von der empirischen Überprüfung der Zuverlässigkeit abgesehen und diese in Übereinstimmung mit einschlägiger Forschungsliteratur (z.B. Hartson et al., 2001; Koutsabasis et al., 2007; Sears, 1997) nicht priorisiert. Da eine hohe Zuverlässigkeit üblicherweise nur durch Standardisierung erreicht werden kann, was wiederum zu einer niedrigen Variabilität bei der Durchführung einer formativen Evaluation und der dabei gefundenen Nutzungsprobleme führen kann, kann dies beispielsweise gleichzeitig die Gründlichkeit reduzieren. Das liegt daran, dass man als Evaluator aufgrund der Standardisierung mit hoher Wahrscheinlichkeit auf die gleichen Nutzungsprobleme wie andere Evaluatoren stößt. In der Praxis, in

der formative Evaluationsmethoden üblicherweise von Expertengruppen angewendet werden, kann dies auch die gesamte Gründlichkeit der Expertengruppe verringern, sofern die individuelle Erkennungsrate eines jeden Evaluators nicht gleichzeitig auch verbessert wird (Hartson et al., 2001). Hierbei kommt meist erschwerend hinzu, dass während der Evaluation oft auch ein Evaluatoreffekt auftritt (z.B. Hertzum & Jacobsen, 2001; Hertzum, Molich, & Jacobsen, 2014; Hornbæk & Frøkjær, 2008; Jacobsen, Hertzum, & John, 1998), weswegen verschiedene Evaluatoren zu unterschiedlichen Nutzungsproblemen bei der Untersuchung desselben Systems kommen können (Hertzum & Jacobsen, 2001). Um damit umzugehen, raten Hartson et al. (2001) die Gründlichkeit und Gültigkeit zunächst immer gegenüber der Zuverlässigkeit zu priorisieren, da für deren Interpretation bereits Erkenntnisse bezüglich der Gründlichkeit und Gültigkeit der Methode hilfreich sein können. Die Zuverlässigkeit, zumindest Gruppenzuverlässigkeit, wird meistens im Zuge des Evaluationsprozesses der Methode besser und wird deswegen in der Praxis meistens nicht explizit betrachtet (Hartson et al., 2001). Zusätzlich raten Nørgaard und Hornbæk (2006) eine detaillierte Datenanalyse durchzuführen, was im Rahmen der folgenden vier Experimente anhand der handlungsorientierten Fehlertaxonomie (Zapf et al., 1989) und eines von Burmester (2016) übernommenen Vorgehens zur Extraktion einzigartiger Nutzungsprobleme realisiert wurde.

Um den Evaluatoreffekt weiter kontrollieren zu können, wurde sich bei den vier Experimenten für das Konstanthalten der Störvariablen *Evaluatorexpertise* und damit für Kontrolle der Sekundärvarianz entschieden (siehe Sarris, 1990). In jedem Experiment führte daher der gleiche Evaluator unter Anleitung des Versuchsleiters (d.h. Verfassers dieser Dissertation) beide Versuchsbedingungen (d.h. unabhängige Variable *Art der formativen Evaluationsmethode*: IntuiBeat-F vs. Nutzertest mit retrospektivem Think-Aloud-Protokoll) durch. Über alle vier Experimente hinweg, kamen insgesamt zwei Studierende des Bachelorstudiengangs Mensch-Computer-Systeme als Evaluatoren zum Einsatz, wobei ein Evaluator die Durchführung des vierten und fünften Experiments (Evaluator 1: männlich, Anfang Zwanzig, 5. Semester) und eine Evaluatorin (Evaluator 2: weiblich, Anfang Zwanzig, 5. Semester) die Durchführung des sechsten und siebten Experiments übernahmen. Beide Evaluatoren verfügten über geringe Erfahrung bei der Durchführung von Nutzertests (d.h. hatten zu diesem Zeitpunkt lediglich einen Nutzertest zuvor durchgeführt), aber brachten eine hohe Kenntnis der jeweiligen Untersuchungsgegenstände mit.

Empirische Überprüfung der Gründlichkeit und Gültigkeit

Beim vierten Experiment (siehe Abschnitt 6.1) kam für die Meta-Evaluation von IntuiBeat-F eine weniger intuitiv benutzbare Software und beim fünften Experiment (siehe Abschnitt 6.2) eine stärker intuitiv benutzbare Software als Untersuchungsgegenstand zum Einsatz. Um zusätzlich feststellen zu können, ob sich die wissenschaftliche Güte von IntuiBeat-F verändert, wenn IntuiBeat-F später von einem Evaluator *strikt* (d.h. IntuiBeat-F wird als verbindliche Unterstützung genutzt: Bei der Überprüfung der beiden Gütekriterien werden ausschließlich kritische Ereignisse für die Ableitung von Nutzungsproblemen berücksichtigt, die mithilfe eines Rhythmus-Peaks entdeckt werden konnten) oder *weniger strikt* (d.h. IntuiBeat-F wird lediglich als unverbindliche Unterstützung genutzt: Bei der Überprüfung der beiden Gütekriterien werden auch kritische Ereignisse für die Ablei-

tung von Nutzungsproblemen berücksichtigt, auch wenn diese nicht explizit mithilfe eines Rhythmus-Peaks entdeckt wurden) angewendet wird, wurden Gültigkeit und Gründlichkeit immer in Abhängigkeit dieses Faktors berichtet und dies bei der Formulierung der Hypothesen entsprechend berücksichtigt.

Da im Zuge des vierten (siehe Abschnitt 6.1) und fünften Experiments (siehe Abschnitt 6.2) das wissenschaftliche Gütekriterium der Gründlichkeit bei der strikten Anwendung von IntuiBeat-F nicht bestätigt werden konnte, was womöglich daran lag, dass der Evaluator wegen der unübersichtlichen Präsentation der Rhythmus-Peaks in Form einer Tabelle einige Rhythmus-Peaks übersehen hatte, kam für das sechste (siehe Abschnitt 6.3) und siebte Experiment (siehe Abschnitt 6.4) eine zusätzliche Analysesoftware für das retrospektive Interview zum Einsatz. Unter Berücksichtigung dieser Analysesoftware wurden im sechsten Experiment eine weniger intuitiv benutzbare Software und im siebten Experiment eine stärker intuitiv benutzbare Software formativ evaluiert. Durch den Einsatz der Analysesoftware konnte das Gütekriterium der Gründlichkeit in beiden Experimenten schließlich auch bei strikter Anwendung bestätigt werden. Anhand der vier Experimente konnte somit die zweite Forschungsfrage und der formative Aspekt der dritten Forschungsfrage beantwortet werden.

Wie bereits im vorangegangenen Kapitel erwähnt (siehe Kapitel 5), ist an dieser Stelle noch darauf hinzuweisen, dass die vier Experimente nicht nur die Beantwortung der beiden Forschungsfragen und damit die Meta-Evaluation von IntuiBeat-F als Ziel hatten. Mithilfe dieser Experimente sollten auch gleichzeitig Nutzungsprobleme bei der Systeminteraktion im Sinne einer formativen Evaluation durch retrospektive Interviews mit den Versuchspersonen identifiziert werden. Wie beim vorangegangenen Kapitel wurden bei den vier Experimenten nicht die dieselben Stimuli (d.h. Softwarepakete) und Evaluatoren verwendet, da kontinuierlich neue qualitative Informationen von diversen CUIs für die Entwicklung von 3D-CUI-Interaktionspatterns benötigt wurden. Die Meta-Evaluation von IntuiBeat-F musste daher dazu parallel laufen. Da die damit verbundenen qualitativen Informationen für die Beantwortung der beiden Forschungsfragen, die die Meta-Evaluation von IntuiBeat-F betreffen, jedoch unerheblich sind, soll auf diese Ergebnisse nur in quantitativer Form mit Bezug zu den wissenschaftlichen Gütekriterien und Forschungsfragen eingegangen werden.

6.1 Experiment 4

Das in diesem Abschnitt vorgestellte Experiment verfolgte das Ziel zu überprüfen, inwiefern IntuiBeat-F wissenschaftliche Güte bezüglich der formalen Hauptgütekriterien *Gründlichkeit* und *Gültigkeit*, sowie *zeitliche Anwendungseffizienz* beim Vergleich mit einem als Quasi-Außenkriterium fungierenden Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der Nutzung der CUI-Regalplanungssoftware *IPO.Rack* (IPO.Plan GmbH, 2015) aufweist, bei der es sich, auf Basis einer Experteneinschätzung, um eine weniger intuitiv benutzbare Software handelt. Im vierten Experiment wurde dementsprechend die zweite Forschungsfrage und der formative Aspekt der dritten Forschungsfrage dieser Arbeit betrachtet.

Die einzelnen Teilabschnitte gehen nun detailliert auf die dem Experiment zugrunde gelegten Hypothesen, die verwendete experimentelle Methode, die Ergebnisse und die Diskussion der Ergebnisse ein.

6.1.1 Überprüfung der Gründlichkeit und Gültigkeit von IntuiBeat-F

Hypothesen

Um IntuiBeat-F wissenschaftliche Güte hinsichtlich Gründlichkeit und Gültigkeit attestieren zu können, müssen diese beiden Gütekriterien bei IntuiBeat-F im Vergleich zum als Quasi-Außenkriterium fungierenden Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der formativen Evaluation der Regalplanungssoftware *IPO.Rack* höher ausgeprägt sein (siehe Abschnitt 3.3), woraus sich die folgenden beiden Hypothesen ableiten lassen:

H1 (Überprüfung der Gründlichkeit) Bei der formativen Evaluation mit IntuiBeat-F sollte sich ein höherer Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der vorhandenen realen Nutzungsprobleme (d.h. Gründlichkeit) als bei dem Nutzertest mit retrospektivem Think-Aloud-Protokoll zeigen.

- H1.A (Weniger Strikt: alle mit IntuiBeat-F abgeleiteten Nutzungsprobleme) Bei der formativen Evaluation mit IntuiBeat-F sollte sich ein höherer Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der vorhandenen realen Nutzungsprobleme (d.h. Gründlichkeit) als bei dem Nutzertest mit retrospektivem Think-Aloud-Protokoll zeigen, wenn alle mit IntuiBeat-F abgeleiteten Nutzungsprobleme berücksichtigt wurden.

- H1.B (Strikt: ausschließlich mit IntuiBeat-F algorithmisch abgeleitete Nutzungsprobleme) Bei der formativen Evaluation mit IntuiBeat-F sollte sich ein höherer Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der vorhandenen realen Nutzungsprobleme (d.h. Gründlichkeit) als bei dem Nutzertest mit retrospektivem Think-Aloud-Protokoll zeigen, wenn ausschließlich auf Basis des Algorithmus von IntuiBeat-F abgeleitete Nutzungsprobleme (d.h. Probleme identifiziert durch Rhythmus-Peaks) berücksichtigt wurden.

H2 (Überprüfung der Gültigkeit) Bei der formativen Evaluation mit IntuiBeat-F sollte sich ein höherer Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der von der jeweiligen Evaluationsmethode gefundenen Nutzungsprobleme (d.h. Gültigkeit) als bei dem Nutzertest mit retrospektivem Think-Aloud-Protokoll zeigen.

- H2.A (Weniger strikt: alle mit IntuiBeat-F abgeleiteten Nutzungsprobleme) Bei der formativen Evaluation mit IntuiBeat-F sollte sich ein höherer Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der von der jeweiligen Evaluationsmethode gefundenen Nutzungsprobleme (d.h. Gültigkeit) als bei dem Nutzertest mit retrospektivem Think-Aloud-Protokoll zeigen, wenn alle mit IntuiBeat-F abgeleiteten Nutzungsprobleme berücksichtigt wurden.

- H2.B (Strikt: ausschließlich mit IntuiBeat-F algorithmisch abgeleitete Nutzungsprobleme) Bei der formativen Evaluation mit IntuiBeat-F sollte sich ein höherer Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der von der jeweiligen Evaluationsmethode gefundenen Nutzungsprobleme (d.h. Gültigkeit)

als bei dem Nutzertest mit retrospektivem Think-Aloud-Protokoll zeigen, wenn ausschließlich auf Basis des Algorithmus von IntuiBeat-F abgeleitete Nutzungsprobleme (d.h. Probleme identifiziert durch Rhythmus-Peaks) berücksichtigt wurden.

6.1.2 Überprüfung der zeitlichen Anwendungseffizienz von IntuiBeat-F

Hypothesen

Um der formativen Evaluationsmethode IntuiBeat-F zeitliche Anwendungseffizienz und damit einen für das Projekt 3D-GUIde wichtigen Teilaspekt praktischer Güte attestieren zu können, muss diese bei IntuiBeat-F im Vergleich zum als Quasi-Außenkriterium fungierenden Nutzertest mit retrospektivem Think-Aloud-Protokoll höher ausgeprägt sein, woraus sich die folgende Hypothese ableiten lässt:

H3 (Zeitliche Anwendungseffizienz von IntuiBeat-F) Bei der formativen Evaluation mit IntuiBeat-F sollte sich eine höhere zeitliche Anwendungseffizienz erkennen lassen, als bei dem Nutzertest mit retrospektivem Think-Aloud-Protokoll.

6.1.3 Methode

6.1.3.1 Teilnehmer

Für das vierte Experiment wurden 25 Versuchspersonen über das Probandensystem des Instituts für Mensch-Computer-Medien an der Universität Würzburg rekrutiert. Da bei einer Person die getestete Software *IPO.Rack* während der Bearbeitung abstürzte, musste dieser Datensatz aus der Datenauswertung ausgeschlossen werden. Demzufolge konnten für die Meta-Evaluation von IntuiBeat-F 24 Versuchspersonen berücksichtigt werden, welche alle rechtsfüßig (d.h. der rechte Fuß stellte den dominanten Fuß dar) waren. Die Versuchspersonen setzten sich dabei aus 19 Frauen und fünf Männern zusammen. Das Durchschnittsalter betrug 20.88 Jahre ($SD = 2.12$). Es handelte sich bei allen Teilnehmern um Studierende der Julius-Maximilians-Universität Würzburg, wobei vier Personen Mensch-Computer-Systeme (16.70 %) und 20 Personen Medienkommunikation (83.30 %) studierten. Alle Versuchsteilnehmer wurden über das Probanden-System des Instituts Mensch-Computer-Medien über eine gesonderte Mail darauf hingewiesen, für den Versuch flache Sportschuhe zu tragen, um eine möglichst problemlose Rhythmuseingabe über das USB-Fußpedal zu ermöglichen. Für die Teilnahme an der Untersuchung bekam jede Versuchsperson eine Versuchspersonenstunde gutgeschrieben. Die mit einem TFQ gemessene Vorerfahrung der Versuchspersonen bezüglich der Nutzung von CUIs betrug im Durchschnitt .07 ($SD = .25$) bei einem Maximum von 6 und lag damit, wie bei der Stichprobe erwartet, im unteren Bereich. Alle Versuchspersonen besaßen damit eine geringe Vorerfahrung mit CUIs. Alle Versuchspersonen gaben an, am Experiment freiwillig teilzunehmen.

6.1.3.2 Versuchsdesign

Für die Beantwortung der zweiten Forschungsfrage und des formativen Aspekts der dritten Forschungsfrage wurde ein einfaktorielles Between-Subjects Design genutzt. Als unabhängige Variable fungierte die *Art der formativen Evaluationsmethode* mit den Ausprägungen: IntuiBeat-F vs. Nutzertest mit retrospektivem Think-Aloud-Protokoll. Als abhängige Variablen fungierten der *Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der vorhandenen realen Nutzungsprobleme* (d.h. *Gründlichkeit*), der *Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der von der jeweiligen Evaluationsmethode gefundenen Nutzungsprobleme* (d.h. *Gültigkeit*) und die *zeitliche Anwendungseffizienz*.

6.1.3.3 Versuchsmaterialien und Maße

Untersuchungsgegenstand der formativen Evaluation: IPO.Rack

Als Untersuchungsgegenstand für die Meta-Evaluation von IntuiBeat-F kam die Regalplanungssoftware *IPO.Rack* der IPO.Plan GmbH (2015) auf Basis einer qualitativen Experteneinschätzung ($N_{Experte}$= 5; Vorgehen: siehe entsprechenden Absatz innerhalb des Teilabschnitts 5.1.4.3) als weniger intuitiv benutzbare Software zum Einsatz. Aufgrund der Tatsache, dass es sich bei einer formativen Evaluation im Gegensatz zu einer summativen Evaluation um eine analysierende Evaluation handelt, die auf eine kontinuierliche Optimierung des Untersuchungsgegenstands anhand von überwiegend qualitativen Daten abzielt, wurden bei der Meta-Evaluation von IntuiBeat-F auch nicht zwei unterschiedlich intuitiv gestaltete Softwareanwendungen verglichen. Es wurde sich stattdessen, so wie es bei einer gestaltend-formativen Evaluation üblich ist (Sarodnick & Brau, 2006), auf die Entdeckung von Nutzungsproblemen und die Ableitung von Verbesserungen bei einem bestimmten System (d.h. Regalplanungssoftware *IPO.Rack*) konzentriert, wobei letztere im Rahmen dieser Arbeit nicht thematisiert werden (siehe Teilabschnitt 3.2.3).

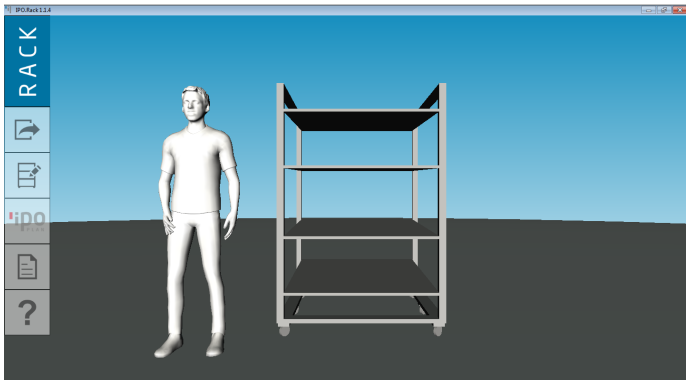


Abbildung 6.1. Die im Rahmen des vierten Experiments als Untersuchungsgegenstand verwendete Regalplanungssoftware *IPO.Rack* der IPO.Plan GmbH (2015).

Die Regalplanungssoftware *IPO.Rack* wurde als Untersuchungsgegenstand für das vierte Experiment gewählt (siehe Abbildung 6.1), da diese Software im Rahmen des Projekts 3D-GUIde evaluiert werden musste (d.h. die IPO.Plan GmbH war ein Industriepartner im Projekt) und man aufgrund des engen Zeitplans im Projekt so viele Synergien wie

möglich nutzen wollte. Obwohl der Einsatzzweck von IPO.Rack mit der Planung von Industrieregalen relativ limitiert ist, war nach einer kurzen Experteneinschätzung dennoch zu erwarten, dass IPO.Rack für die im Experiment verwendete Studierendenstichprobe eine eher weniger intuitiv benutzbare Softwareanwendung darstellen würde (d.h. weniger intuitiv benutzbare Softwareanwendungen sollten logischerweise auch zu vielen Nutzungsproblemen führen). IPO.Rack würde sich daher gut als Untersuchungsgegenstand eignen, da damit die miteinander verglichene formative Evaluationsmethoden (d.h. IntuiBeat-F vs. Nutzertest mit retrospektivem Think-Aloud-Protokoll) ihr Potential zur Entdeckung von Nutzungsproblemen zeigen können.

Durch die selben Experten wurden in Abstimmung mit der IPO.Plan GmbH fünf experimentelle Aufgaben gewählt (siehe Anhang B.4.1), die die Funktionalität von IPO.Rack so gut wie möglich repräsentieren sollten. Sie wurden von den Versuchspersonen in einer festen Reihenfolge bearbeitet. Bei den gewählten Aufgaben wurde zusätzlich darauf geachtet, dass die von LaViola et al. (2017) beschriebenen domänenübergreifenden, grundlegenden Nutzeraufgaben Selektion, Manipulation, Navigation, Zeicheneingaben und Systemsteuerung von diesen abgedeckt waren, um wie beim ersten Experiment (siehe Teilabschnitt 5.1.4.3) und den anderen vorigen Experimenten intuitiv benutzbare 3D-CUI-Interaktionslösungen künftig als Interaktionspatterns auf Basis der Ergebnisse ableiten zu können. Alle Aufgaben wurden dabei in ein fiktives Anlagenplanungsszenario eines Industrieregals bei einem Autohersteller eingebettet.

Hierbei mussten die Versuchspersonen zunächst einen bestimmten Regaltyp und Regalhersteller des zu planenden Regals festlegen (erste Aufgabe). Im Anschluss mussten die Versuchspersonen die Anzahl und Dimensionen (d.h. Höhe, Anstellwinkel und Breite) der Regalböden, sowie Höhe, Breite und Tiefe des gesamten Regals basierend auf den ihnen zur Verfügung gestellten Angaben spezifizieren (zweite Aufgabe). Daraufhin mussten die Versuchspersonen die Regalböden und das gesamte Regal in bestimmten Farben einfärben (dritte Aufgabe). Danach mussten die Versuchspersonen die Stückliste für das geplante Regal, damit dieses so montierbar war, als PDF auf dem Desktop ihres PCs (d.h. Versuchspersonen-PC) bereitstellen. Als letzte Aufgabe mussten die Versuchspersonen eine nicht selbst erstellte Regalkonfiguration (d.h. Sammlung mehrerer geplanter Regale) namens „AutoRegale.dat“ laden und davon ein bestimmtes Regal anpassen. Die Versuchspersonen mussten bei diesem Regal jedem Regalboden zwei Rollbahnen hinzufügen, damit später darauf Stoßstangen gelagert werden können. Nachdem sie Rollbahnen hinzugefügt hatten, mussten sie das angepasste Regal innerhalb einer neuen Regalkonfiguration abspeichern (fünfte Aufgabe).

Formative Evaluation intuitiver Benutzung mit IntuiBeat-F

Wie bereits an verschiedenen Stellen dieser Arbeit erwähnt, verbirgt sich hinter der Bezeichnung *IntuiBeat-F* eine Adaption der Rhythmusmethode (siehe Korbach et al., 2018; Park & Brünken, 2015) und der regelmäßigen Klopfaufgabe von M. J. Albers (2011) für die formative Evaluation intuitiver Benutzung. Da erstere noch nicht zur formativen Evaluation im Rahmen eines Nutzertests eingesetzt wurde und die wissenschaftliche Güte von letzterer aufgrund gravierender methodischer Limitationen (d.h. (1) keine erhobene

individuelle Baseline der Versuchsteilnehmer, (2) keine Inhibitionsaufgabe und (3) die ausschließlich von den musikalischen Fähigkeiten des Evaluators abhängenden Identifikation potentieller Nutzungsprobleme) anzuzweifeln ist (siehe Abschnitt 4.2), wurde für die Adaption *IntuiBeat-F* die Aufzeichnung und Auswertung von Rhythmusabweichungen nach dem Vorbild der Rhythmusmethode (siehe Korbach et al., 2018; Park & Brünken, 2015) von IntuiBeat-S übernommen (siehe Kapitel 5) und zusätzlich noch ein Algorithmus (d.h. AMPD: Automatic multi scale-based peak detection) von Scholkmann, Boss und Wolf (2012) für die eindeutige Identifikation kritischer Ereignisse genutzt. Mithilfe dieses Algorithmus sollte der Willkürlichkeit bei der Identifikation potentieller Nutzungsprobleme durch den Evaluator entgegengewirkt, und stattdessen dem Evaluator die Spitzen (d.h. Rhythmus-Peaks) der Rhythmusabweichungen als „sichere“ Hinweise auf potentielle Nutzungsprobleme (d.h. kritische Ereignisse) bereitgestellt werden. Der AMPD wurde zwar noch nicht zuvor für diesen Zweck eingesetzt, aber dennoch für die Identifikation lokaler Extrema (d.h. Rhythmus-Peaks) als geeigneter Algorithmus erachtet, da rauschfreie Daten bei der Durchführung eines Nutzertests außerhalb einer experimentellen Idealsituation nicht zu erwarten sind, und er bereits seine Effektivität bei der Identifikation von Spitzen in quasi-periodischen Signalen in anderen Domänen demonstrieren konnte (siehe Scholkmann et al., 2012). Für die genaue Funktionsweise und mathematischen Hintergründe des Algorithmus sei an dieser Stelle auf Scholkmann et al. (2012) verwiesen.

Im Folgenden soll nun die Vorgehensweise von IntuiBeat-F zur formativen Evaluation intuitiver Benutzung beim vierten Experiment beschrieben werden. Diese Beschreibungen lassen sich jedoch auch auf andere Untersuchungsgegenstände generalisieren und fungieren damit als eine generelle Methodenbeschreibung von IntuiBeat-F.

Hard- und Software von IntuiBeat-F

Da in der Praxis bei einer formativen Evaluation üblicherweise verschiedene Arten von Daten (z.B. auch quantitative Daten von Leistungsmaßen, die eigentlich zur summativen Evaluation eingesetzt werden) zum Auffinden von Nutzungsproblemen zusammengeführt werden (d.h. Triangulation, siehe Dumas et al., 1999), es zur Minimierung der Entwicklungszeiten bei einem Industrieprojekt Usus ist, bei einer Evaluation unabhängig vom Evaluationsziel so viele Informationen wie möglich zu sammeln und damit die beiden prinzipiellen Evaluationsarten miteinander zu kombinieren (Burmester, 2016), wurde sich entschieden, die für IntuiBeat-F benötigte Funktionalität in die bestehende für IntuiBeat-S genutzte IntuiBeat-Software zu integrieren. Die Nutzung der IntuiBeat-Software ist bei der Anwendung von IntuiBeat-F dementsprechend mit der Anwendung von IntuiBeat-S nahezu identisch, weswegen in diesem Absatz gemäß nur auf drei Unterschiede bezüglich des Experimentalmodus der IntuiBeat-Software eingegangen wird. Für die allgemeine Beschreibung der Anwendung der IntuiBeat-Software sei an dieser Stelle auf den entsprechenden Abschnitt des ersten Experiments (siehe Abschnitt 5.1.4.3) verwiesen.

Funktionsweise des Experimental-Modus der IntuiBeat-Software bei der Anwendung von IntuiBeat-F und Beschreibung des Programmablaufs

Zur formativen Evaluation intuitiver Benutzung von IPO.Rack führte der Evaluator die Experimentalmessung wie bei IntuiBeat-S mit drei Unterschieden durch:

1. Nachdem der Evaluator die Rhythmusaufzeichnung im Experimentalmodus gestoppt hatte, zeigte die Software nicht nur die durchschnittlichen Baseline-Abweichungen innerhalb des kurzen und des langen Rhythmusintervalls an, sondern auch eine Liste der mithilfe des AMPD-Algorithmus ermittelten Rhythmus-Peaks (d.h. kritische Ereignisse).
2. Nachdem der Evaluator die Rhythmusaufzeichnung gespeichert hatte, wurden anstelle von drei verschiedenen Logs, vier verschiedene Logs als CSV-Dateien (d.h. Typ „[ID]_raw.csv“, Typ „[ID]_clean.csv“, Typ „[ID]_infos.csv“, Typ „[ID]_marker.csv“) generiert. Die im Vergleich zur Experimentalmessung bei IntuiBeat-S (siehe Absatz 5.1.4.3 des ersten Experiments) neue Marker-Datei enthielt die mithilfe des AMPD-Algorithmus ermittelten Ausprägungen der Rhythmus-Peaks (kurzes Rhythmusintervall), einen Zeitstempel in Millisekunden (Systemzeit) und die vergangene Zeit der Rhythmusaufzeichnung bis zum jeweiligen Rhythmus-Peak in Sekunden.
3. Während der eigentlichen Aufgabenbearbeitung erfolgte keinerlei Beobachtung der Versuchspersonen und damit auch keine Protokollierung von kritischen Ereignissen, die auf potentielle Nutzungsprobleme hinweisen konnten. Es wurde sich gegen ein derartiges Vorgehen entschieden, obwohl es bei der Anwendung eines Think-Aloud-Protokolls im Rahmen eines Nutzertests üblich ist (siehe Nørgaard & Hornbæk, 2006; Sarodnick & Brau, 2006; Van den Haak & De Jong, 2003). Da eine Reihe von Studien die Abhängigkeit des Erfolgs einer überwiegend qualitativen, formativen Evaluation vom Expertenwissen des Evaluators als potentielle Gefahr zeigen (Saroyan, 1992; Scriven, 1996), sollte durch dieses Vorgehen die Meta-Evaluation von IntuiBeat-F möglichst unabhängig vom Expertenwissen des Evaluators erfolgen können, um so den Evaluatoreffekt zu reduzieren (siehe Alshamari & Mayhew, 2009; Nørgaard & Hornbæk, 2006). Darüber hinaus konnten Studien (siehe Akers, 2010) zur Methodik der kritischen Ereignisse (Del Galdo, Williges, Williges, & Wixon, 1986; Flanagan, 1954; Hartson & Castillo, 1998), einer Vorgehensweise (d.h. entscheidende positive und negative Ereignisse werden während der Systeminteraktion von einem geschulten Beobachter festgehalten), die einem jeden formativen Nutzertest zugrunde liegt (Capra, 2002), zeigen, dass eine besonders hohe zeitliche Anwendungseffizienz bei einer formativen Evaluation dann zu beobachten ist, wenn die Ereignisse von den Testteilnehmern selbst, ohne große Beteiligung des Evaluators, berichtet werden. Akers (2010) konnte eine hohe zeitliche Anwendungseffizienz auch im Bereich von CUIs für eine, auf Selbstauskunft fokussierte, formative Evaluationsmethode bestätigen. Dementsprechend wurde die Identifikation von Nutzungsproblemen vollständig auf das retrospektive Interview verlagert, in dem der Evaluator die Rhythmus-Peaks, die ihm am Ende der Aufzeichnung durch die IntuiBeat-Software angezeigt und in einer Marker-Datei (geöffnet in einem Tabellenkalkulationsprogramm) bereitgestellt wurden, als „sichere“ Anhaltspunkte zur Identifikation von kritischen Ereignissen nutzen konnte.

Nachdem der Evaluator die Experimentalmessung unter Berücksichtigung der oben aufgeführten drei Unterschiede für alle experimentellen Aufgaben mithilfe von IntuiBeat-F durchgeführt hatte, führte dieser mit den Versuchspersonen ein retrospektives Interview bezüglich aller Aufgaben durch. Er spielte den Versuchspersonen dazu die Bildschirmaufzeichnung ihrer Aufgabenbearbeitung nacheinander auf seinem Evaluatoren-PC (siehe Absatz „Apparatur“) vor. Dabei instruierte er sie zuvor, währenddessen ihre mit der Aufgabenbearbeitung einhergehenden Gedanken zu verbalisieren und dementsprechend ein retrospektives Think-Aloud-Protokoll anzuwenden. Im Rahmen dieser standardisierten mündlichen Instruktion wurde den Versuchspersonen auch mitgeteilt, dass es das Ziel dieses retrospektiven Interviews sei, Schwachstellen bei der Benutzung, also Nutzungsprobleme aufzudecken und mit der Hilfe der Versuchspersonen konkrete Verbesserungsvorschläge zu entwickeln. Es wurde im Rahmen der mündlichen Instruktion vom Evaluator besonders darauf hingewiesen, dass es nicht um die Beurteilung der Versuchspersonen ging, sondern lediglich um die Beurteilung der Gestaltung der getesteten Softwareanwendung.

Das retrospektive Think-Aloud-Protokoll wurde bei IntuiBeat-F dem parallelen, während der eigentlichen Aufgabenbearbeitung durchgeführten, Think-Aloud-Protokoll vorgezogen, da, wie bereits in Abschnitt 3.2.3 erläutert, das retrospektive Interview weniger Intrusion bei der Aufgabenbearbeitung erzeugt und damit bei einer retrospektiven Verbalisierung mehr echte Nutzungsprobleme identifiziert werden können (siehe V. A. Bowers & Snyder, 1990; Van Den Haak et al., 2003; Van den Haak et al., 2004). Der Evaluator öffnete im Anschluss an die Instruktion die Bildschirmaufzeichnung der ersten Aufgabe auf dem Evaluatoren-PC und die entsprechende für das retrospektive Interview benötigte Marker-Datei auf dem Analyse-PC (siehe Absatz „Apparatur“ dieses Experiments) in einem Tabellenkalkulationsprogramm. Er bat daraufhin die Versuchspersonen ihre Aufgabenbearbeitung parallel zur gezeigten Bildschirmaufzeichnung zu kommentieren. Währenddessen protokollierte der Evaluator auf Basis der Äußerungen der Versuchsteilnehmer und seiner Beobachtungen des Videomaterials in Anlehnung an Burmester (2016) kritische Ereignisse, die auf potentielle Nutzungsprobleme der Versuchspersonen hinwiesen, jedoch ohne interpretative Anreicherungen. Bei Unklarheiten stoppte er die Videoaufzeichnung an der entsprechenden Stelle, um diese mit den Versuchsteilnehmern zu klären. Der Evaluator stoppte außerdem an den in der geöffneten Marker-Datei enthaltenen Rhythmus-Peaks in der Videoaufzeichnung, um zusammen mit den Versuchspersonen herauszufinden, welches kritische Ereignis eine Spitze bei der Rhythmusaufzeichnung verursacht hat, und ob dahinter ein potentielles Nutzungsproblem steckt. Aufgrund der Tatsache, dass ein solches Ereignis immer notwendigerweise vor einem erhöhte mentale Beanspruchung indizierenden Rhythmus-Peak stattgefunden haben muss, sprang der Evaluator drei Sekunden vor den in der Marker-Datei enthaltenen Zeitpunkt.

Wie bereits in Kapitel 4 erläutert, konnten Park und Brünken (2015) lediglich beim kurzen Rhythmusintervall einen Zusammenhang mit der mentalen Beanspruchung des Nutzers feststellen, was sie darauf zurückführten, dass die Inhibition des Rhythmus bei der Aufgabenbearbeitung auf ein zeitlich vorgelagertes kritisches Ereignis hinweist (d.h. Rhythmus-Peak entsteht erst nach mental anstrengendem Ereignis und nicht zeitgleich), ein Phänomen, was in Form von ereigniskorrelierten Potentialen sehr gut erforscht ist (z.B. Berti, 2008; Kiefer, Marzinzik, Weisbrod, Scherg, & Spitzer, 1998; Rugg & Coles, 1995). Da die Zeitspanne zwischen Ereignis und Rhythmus-Peak aufgrund der unterschiedlichen Komplexität bei der kognitiven Informationsverarbeitung, die zu dem kritischen Ereignis geführt

hat, nicht bestimmbar ist (d.h. es kann kein universeller Wert für alle kritischen Ereignisse festgelegt werden), wurde stattdessen drei Sekunden vor und nach dem Rhythmus-Peak mit Unterstützung der Versuchspersonen nach dem für den Rhythmus-Peak verantwortlichen Ereignis gesucht.

Laut Reinhardt et al. (2018) dauert eine Interaktion unter Berücksichtigung der Technik des Keystroke Level Modeling (siehe Card et al., 1980) bei einem Desktop User Interface wie einem CUI nicht länger als drei Sekunden, weswegen man im Umfeld von drei Sekunden um den Rhythmus-Peak leicht die für das Ereignis verantwortliche Interaktion und das dazugehörige Steuerelement ausfindig machen kann, ohne dabei groß an zeitlicher Anwendungseffizienz zu verlieren. Der Evaluator protokollierte jedes mithilfe des beschriebenen Vorgehens identifizierte Ereignis handschriftlich. Hierbei wurde insbesondere das Ereignis selbst (d.h. kurze Beschreibung des Ereignisses), der zum Ereignis gehörende Zeitstempel in der analysierten Videoaufzeichnung (Systemzeit), das vom Ereignis betroffene Steuerelement und inwiefern, ein Rhythmus-Peak den Evaluator bei der Entdeckung des Ereignisses unterstützt hat oder nicht, festgehalten. Letzter Aspekt war wichtig, weil im Zuge der Bestimmung der Gründlichkeit und Gültigkeit ja zwischen weniger strikten (d.h. alle kritischen Ereignisse fließen in die Ableitung von Nutzungsproblemen ein) und strikten (d.h. nur kritische Ereignisse mit Rhythmus-Peak fließen in die Ableitung von Nutzungsproblemen ein) Anwendung von IntuiBeat-F differenziert werden sollte. Außerdem wurden vom Evaluator zusätzliche interessante Informationen als Kommentar protokolliert, um die formative Evaluation des Untersuchungsgegenstands möglichst praxisnah zu gestalten (siehe Van den Haak & De Jong, 2003). Nachdem der Evaluator mit den Versuchspersonen die Bildschirmaufzeichnung der ersten Aufgabe durchgesprochen hatte, sprach er die übrigen Aufgaben in der gleichen Form durch und notierte sich auf potentielle Nutzungsprobleme hindeutende Ereignisse auf dieselbe Art und Weise auf Papier.

Abschließend soll an dieser Stelle noch angemerkt werden, dass es für die Zuordnung von Rhythmus-Peaks zu den entsprechenden Zeitpunkten der Bildschirmaufzeichnung unabdingbar ist, dass der Evaluator die Bildschirm- und Rhythmusaufzeichnung zum exakt gleichen Zeitpunkt startet, da nur so die Identifikation des für die Erhöhung der mentalen Beanspruchung verantwortlichen Steuerelements bzw. der damit verbundenen Interaktion eindeutig ist. Des Weiteren wurde sich bewusst dagegen entschieden, sich beim retrospektiven Interview nur auf die Rhythmus-Peaks und die damit verbundenen kritischen Ereignisse zu konzentrieren (d.h. Evaluator springt lediglich direkt an die entsprechenden Videostellen ohne den Versuchspersonen die komplette Aufgabenbearbeitung vorzuspielen), da die Erinnerung an bestimmte kritische Ereignisse durch den Kontextbezug (d.h. episodisches Gedächtnis) während dem retrospektiven Interview bedingt wird und es ohne diesen Bezug zu Schwierigkeiten bei der Rekonstruktion von kritischen Ereignissen kommen könnte (siehe Rubin & Umanath, 2015; Rudy, 2009; D. M. Smith & Mizumori, 2006; Tulving, 1985).

Formative Evaluation intuitiver Benutzung durch Nutzertest mit retrospektivem Think-Aloud-Protokoll

Um die wissenschaftlichen Gütekriterien *Gründlichkeit* und *Gültigkeit* von IntuiBeat-F überprüfen zu können, kam neben IntuiBeat-F noch ein klassischer Nutzertest mit retro-

spektivem Think-Aloud-Protokoll zur Identifikation von Nutzungsproblemen zum Einsatz (siehe Abschnitt 3.5). Der Ablauf und die Protokollierung unterschied sich hier im Vergleich zu IntuiBeat-F, wie es im letzten Absatz beschrieben wurde, nur in zwei Punkten. Zum einen hatte der Evaluator beim Nutzertest mit retrospektivem Think-Aloud-Protokoll keine Rhythmus-Peaks zur Verfügung, weswegen er bei der Analyse des Videomaterials keine zusätzliche Unterstützung erhielt. Des Weiteren protokollierte der Evaluator kritische Ereignisse während der eigentlichen Systemnutzung mit, so wie dies bei einem Nutzertests mit retrospektivem Think-Aloud-Protokoll Usus ist, aber auch mit den im letzten Absatz genannten Problemen verbunden ist.

Bewertung der Gründlichkeit und Gültigkeit

Um im Rahmen des vierten Experiments die wissenschaftliche Güte von IntuiBeat-F als formative Methode zur Evaluation intuitiver Benutzung anhand der beiden Gütekriterien *Gründlichkeit* und *Gültigkeit* überprüfen zu können, mussten die, mit den beiden formativen Evaluationsmethoden protokollierten, kritischen Ereignisse zunächst zu einzigartigen Nutzungsproblemen zusammengefasst werden. Hierzu wurden die in Abbildung 6.2 dargestellten Arbeitsschritte Segmentierung, Indexierung und Interpretation in Anlehnung an das Vorgehen von Burmester (2016) verwendet, welches auf Überlegungen von Hassenzahl und Burmester (1999) fußt und sich an qualitative Auswertungsverfahren wie die Globalauswertung von Legewie (1994), offenes und axiales Kodieren der Grounded Theory (Strauss, Corbin, Niewiarra, & Legewie, 1996) und Prinzipien der Fokusgruppenanalyse (Wennecker, 2001) anlehnt.

Im Zuge der Segmentierung wurden zunächst die mit den beiden formativen Evaluationsmethoden handschriftlich protokollierten Ereignisse jeweils in eine chronologische Reihenfolge gebracht und als sogenannte Segmente (z.B. in einem digitalen CSV-Format) festgehalten (siehe „Segmentierung“ in Abbildung 6.2). Im Anschluss erfolgte die Indexierung aller Segmente einzeln für jede der beiden formativen Evaluationsmethoden, bei der jedes Segment analysiert und eindeutig benannt wurde (siehe „Indizes“ in Abbildung 6.2). Hierzu wurden die Indizes (z.B. Problem bei der Auswahl eines Regalherstellers) induktiv aus den Segmenten gebildet (d.h. Evaluator startete mit dem ersten Segment und bildete sofort Indizes) und als neue Tabellenspalte den einzelnen Segmenten in der entsprechenden CSV-Datei hinzugefügt. Die gebildeten Indizes wurden im Laufe der Indexierung iterativ vom Evaluator präzisiert, da es nach einer gewissen Menge von gebildeten Indizes nötig sein kann, einen Index genauer zu fassen oder beispielsweise in zwei Indizes aufzuspalten (siehe Tabelle 6.1). Der Arbeitsschritt der Indexierung bietet den Vorteil, dass man einem Segment zur inhaltlichen Klassifizierung mehrere Indizes zuordnen und es so als Ganzes erhalten kann (Burmester, 2016). Nachdem sich ein Index stabilisiert hatte, verfasste der Evaluator für diesen eine kurze Definition in einer weiteren CSV-Datei. Abschließend ging der Evaluator die stabilen Indexdefinitionen durch und prüfte die Definitionen auf Sinnhaftigkeit, womit die Indexierung laut Burmester (2016) abgeschlossen war.

Tabelle 6.1. *Beispiel für ein Segment mit zwei Indizes, welches im Rahmen des vierten Experiments gebildet wurde. Anmerkung: Der Zeitpunkt innerhalb der Bildschirmaufzeichnung (d.h. Videostelle), wo das kritische Ereignis stattfand, wurde als Systemzeit angegeben, um Rhythmus-Peak und Videoaufzeichnung darüber einfacher synchronisieren zu können.*

Nr.	VP	Aufgabe	Screen	UI-Element	Ereignis	Index	Videostelle	Peak
166	18	1	Verwaltung	Buttons	Den Beschriftungen der Buttons und den Tooltips mangelt es an Aussagekraft.	a) Wording der Tooltips nicht verständlich. b) Wording der Buttons nicht verständlich.	10:26:58 - 10:27:37	1

Laut Burmester (2016) stellt die Zuordnung des Inhalts eines Segments zu Indizes und die damit verbundene Reduzierung des Inhalts eines Segments auf das Wesentliche bereits die erste Stufe der Interpretation, also der Identifikation und der genauen Beschreibung der Nutzungsprobleme dar (siehe „Usability-Probleme“ in Abbildung 6.2). Alle mit den beiden formativen Evaluationsmethoden identifizierten Segmente, die dem gleichen Index zugeordnet wurden, wurden im Anschluss vom Evaluator nacheinander betrachtet und interpretiert, welches Nutzungsproblem durch das betrachtete Segment beschrieben ist. Im Anschluss wurden daraus entsprechende Nutzungsprobleme vom Evaluator abgeleitet und in einer neuen CSV-Datei gespeichert. Bei der Formulierung der Nutzungsprobleme wurde vom Evaluator stets darauf geachtet, dass auch andere Personen (z.B. weitere Mitarbeiter des Projekts 3D-GUIde) daraus Verbesserungen für 3D-CUI-Interaktionslösungen für die Gestaltung von Interaktionspatterns ableiten konnten. Wie bereits in der Einleitung dieser Arbeit erwähnt, spielen diese rein qualitativen Gestaltungsempfehlungen für die Meta-Evaluation von IntuiBeat-F und IntuiBeat-S keine Rolle und werden deswegen auch nicht weiter thematisiert. Dieser Arbeitsschritt ist demzufolge auch nicht in Abbildung 6.2 enthalten.

Nachdem der Evaluator für beide verglichenen formativen Evaluationsmethoden eine vollständige Liste von Nutzungsproblemen abgeleitet hatte, nutzte er im Anschluss die in Teilabschnitt 3.2.3 vorgestellte handlungsorientierte Fehlertaxonomie, um über die Echtheit eines Nutzungsproblems (d.h. „Beeinträchtigt das identifizierte Nutzungsproblem wirklich die intuitive Handlungsregulation und lässt sich dementsprechend in eine der Fehlerkategorien einordnen?“) und über die konkrete Ursache des Nutzungsproblems (d.h. „In welcher Phase der Handlungsregulation und auf welcher Ebene der Informationsverarbeitung fand das Problem statt?“) zu entscheiden (siehe Zapf et al., 1989). Für die Kategorisierung anhand der handlungsorientierten Fehlertaxonomie nutzte der Evaluator einen von Zandbergen (2015) vorgeschlagenen Entscheidungsbaum (siehe Abbildung 6.3), welcher sich wiederum an einem Entscheidungsbaum von Haar (2013) orientierte, der ursprünglich für eine zeitlich effiziente Kategorisierung von Nutzungsproblemen anhand der Taxonomie von Rasmussen (1983) genutzt wurde.

Der Evaluator hatte für die Kategorisierung der Nutzungsprobleme neben dem Entscheidungsbaum auch noch englischsprachige Beschreibungen der einzelnen Fehlerkategorien mit Beispielen ausgedruckt vor sich liegen, die von Zandbergen (2015) ebenfalls ohne Änderungen übernommen wurden. Zandbergen (2015) stellte darüber hinaus noch passende Fragen für die einzelnen Entscheidungen des Baums bereit, um einen Evaluator bei

der Klassifikation noch weiter zu unterstützen. Diese Fragen und eine schrittweise Anleitung zur Benutzung des Baums wurden ebenfalls aus der Arbeit von Zandbergen (2015) übernommen und dem Evaluator für die Klassifikation auch ausgedruckt, in englischer Sprache bereitgestellt. Um die Zuverlässigkeit der Einteilung in die Fehlerkategorien bewerten zu können, wurde ein zweiter Evaluator (d.h. Autor dieser Arbeit) hinzugezogen und die Übereinstimmung der Kodierungen mithilfe von Cohens κ bestimmt. Für die finale Zuordnung diskutierten die beiden Evaluatoren ihre Zuordnungen und einigten sich pro Nutzungsproblem auf eine Ursachenkategorie.

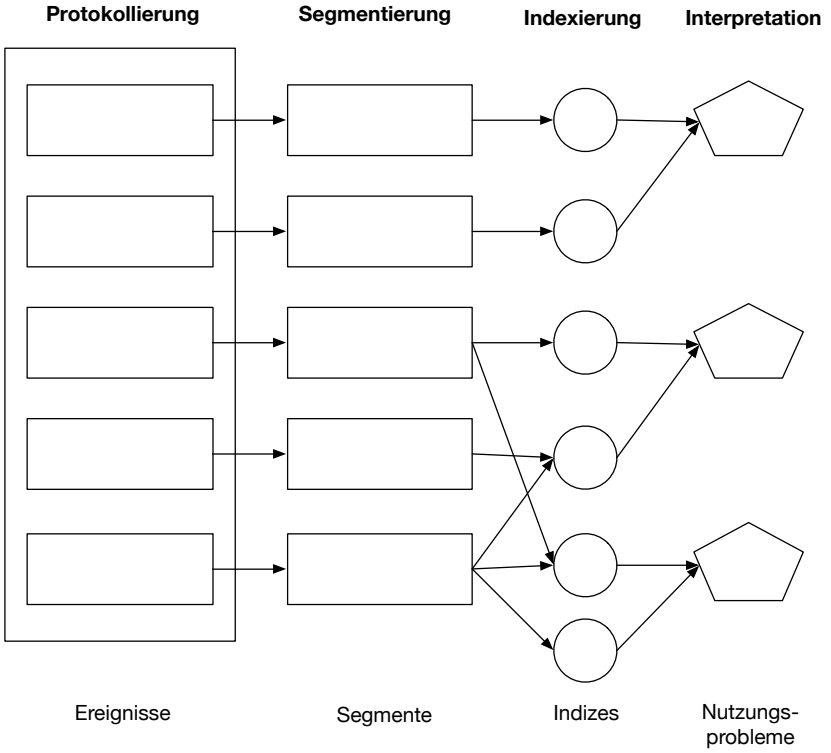


Abbildung 6.2. Arbeitsschritte zur Ableitung von Nutzungsproblemen aus protokollierten kritischen Ereignissen in Anlehnung an Burmester (2016).

Die mithilfe dieses Vorgehens identifizierten Fehlerkategorien wurden vom Evaluator den abgeleiteten Nutzungsproblemen zugeordnet und von ihm in einer weiteren Spalte in der entsprechenden CSV-Datei mit den Nutzungsproblemen vermerkt. Durch diese äußerst detaillierte Form der Datenanalyse wurde zusätzlich der Evaluatoreffekt minimiert, da der von Hertzum und Jacobsen (2001) angesprochene hierfür verantwortliche Hauptfaktor, die großen interindividuellen Unterschiede in der Interpretation kritischer Ereignisse, durch das Vorgehen von Burmester (2016) und die handlungsorientierte Fehlertaxonomie (Zapf et al., 1989) durch Reduktion des Interpretationsspielraums mit hoher Wahrscheinlichkeit verringert werden konnte.

Um die wissenschaftliche Güte von IntuiBeat-F anhand der Gütekriterien *Gründlichkeit* und *Gültigkeit* bewerten zu können, wurde zunächst für jede der beiden Evaluationsmethoden jeweils die Anzahl der gefundenen realen Nutzungsprobleme (d.h. alle Nutzungsprobleme, denen mit der handlungsorientierten Fehlertaxonomie eine Fehlerkategorie abgesehen

vom Bewegungsfehler zugeordnet werden konnte, da intuitive Benutzung nicht durch motorische Ineffizienzen oder zeitliche Ineffizienzen gehemmt wird. Bewegungsfehler stellen daher keine realen Nutzungsprobleme dar, siehe Teilabschnitt 3.2.3) und die Anzahl aller gefundenen Nutzungsprobleme (d.h. alle Nutzungsprobleme, auch Nutzungsprobleme denen mithilfe der handlungsorientierten Fehlertaxonomie ein Bewegungsfehler als Ursache zugeordnet werden konnte) für jede getestete Methode einzeln bestimmt.

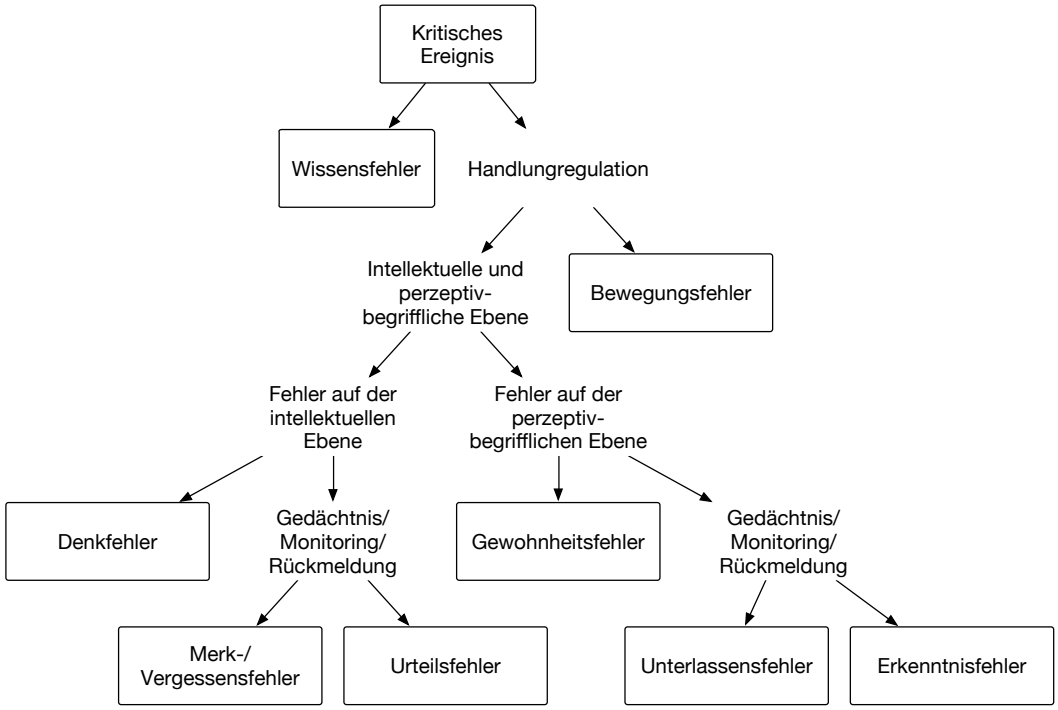


Abbildung 6.3. Entscheidungsbaum zur Kategorisierung von kritischen Ereignissen auf Basis der handlungsorientierten Fehlertaxonomie von Zapf, Brodbeck und Prümper (1989) in Anlehnung an Zandbergen (2015).

Daraufhin konnte in Anlehnung an Hartson et al. (2001), Sears (1997) und Koutsabasis et al. (2007) die Gültigkeit der formativen Evaluationsmethode bezüglich IPO.Rack als Anteil der, durch diese Methode gefundenen, realen Nutzungsprobleme an der Anzahl der, von dieser Methode gefundenen, Nutzungsprobleme durch Teilung der Anzahl der, von dieser Methode gefundenen, Nutzungsprobleme und der Anzahl der, von dieser Methode gefundenen, realen Nutzungsprobleme berechnet und so die abhängige Variable *Gültigkeit* operationalisiert werden (siehe Formel 6.1).

$$Gültigkeit_{Methode} = \frac{\text{Anzahl mit Methode gefundenen realen NPs}}{\text{Gesamtanzahl mit Methode gefundenen NPs}} \tag{6.1}$$

Unterschiede bei der Berechnung der Gültigkeit bei strikter und wenig strikter Anwendung von IntuiBeat-F Bei der Operationalisierung wurde, wie in den Hypothesen formuliert, unterschieden, ob im Falle von IntuiBeat-F für die Berechnung der Gültigkeit alle (d.h. weniger strikt) mit der Methode identifizierten Nutzungsprobleme (AV_Gültigkeit

_WenigerStrikt) oder nur die auf Basis von Rhythmus-Peaks (d.h. strikt) abgeleiteten Nutzungsprobleme (AV_Gültigkeit_Strikt) genutzt wurden. Die abhängige Variable *Gültigkeit* wurde dementsprechend unterschiedlich operationalisiert. Eine geringe Gültigkeit wird dabei durch einen geringen Anteil an gefundenen realen Nutzungsproblemen und eine hohe Gültigkeit durch einen hohen Anteil an gefundenen realen Nutzungsproblemen indiziert (Hartson et al., 2001; Koutsabasis et al., 2007; Sears, 1997). Im Falle des Nutzertests mit retrospektivem Think-Aloud-Protokoll wurde, wie bereits erwähnt, eine derartige Unterscheidung nicht vorgenommen und immer alle identifizierten Nutzungsprobleme zur Berechnung der Gültigkeit verwendet.

Für die Berechnung der Gründlichkeit, welche den Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der vorhandenen realen Nutzungsprobleme in IPO.Rack repräsentiert, musste zunächst die Gesamtzahl vor Berechnung des Anteils bestimmt werden. Zur Operationalisierung der Gesamtzahl der vorhandenen realen Nutzungsprobleme in IPO.Rack wurden die mit beiden formativen Evaluationsmethoden erzeugten Mengen bzw. Listen von realen Nutzungsproblemen in Anlehnung an Hartson et al. (2001), Sears (1997) und Koutsabasis et al. (2007) als Vereinigungsmenge aggregiert. Bei der Vereinigung wurde darauf geachtet, in der aus der Vereinigung resultierenden Gesamtliste keine Duplikate zu haben, weswegen auch hier ein induktiver Ansatz verfolgt wurde. Es wurde sich dementsprechend bei jedem, durch eine der beiden Evaluationsmethoden identifizierten, realen Nutzungsproblem gefragt, inwiefern dieses bereits von der anderen formativen Evaluationsmethode identifiziert wurde, und dementsprechend das gleiche Nutzungsproblem repräsentiert.

Nachdem so eine Gesamtliste für die getestete Softwareanwendung *IPO.Rack* als Approximation für die in Wirklichkeit vorhandene Gesamtzahl an realen Nutzungsproblemen gebildet wurde, konnte anhand dieser Liste die Gesamtzahl der vorhandenen realen Nutzungsproblem bestimmt und durch Teilung dieser Gesamtzahl durch die entsprechende Anzahl der realen gefundenen Nutzungsprobleme, der Anteil von gefundenen realen Nutzungsproblemen für die jeweilige Methode berechnet und so die abhängige Variable *Gründlichkeit* operationalisiert werden (siehe Formel 6.2).

$$Gründlichkeit_{Methode} = \frac{\text{Anzahl mit Methode gefundenen realen NPs}}{\text{Gesamtanzahl aller existierenden realen NPs}} \tag{6.2}$$

Unterschiede bei der Berechnung der Gründlichkeit bei strikter und wenig strikter Anwendung von IntuiBeat-F Bei der Operationalisierung wurde, wie in den Hypothesen formuliert, unterschieden, ob im Falle von IntuiBeat-F für die Berechnung der Gründlichkeit alle (d.h. weniger strikt) mit der Methode identifizierten Nutzungsprobleme (AV_Gründlichkeit_WenigerStrikt) oder nur die auf Basis von Rhythmus-Peaks (d.h. strikt) abgeleiteten Nutzungsprobleme (AV_Gründlichkeit_Strikt) genutzt wurden. Die abhängige Variable *Gründlichkeit* wurde dementsprechend unterschiedlich operationalisiert. Eine geringe Gründlichkeit wird dabei durch einen geringen Anteil an gefundenen realen Nutzungsproblemen und eine hohe Gründlichkeit durch einen hohen Anteil an gefundenen realen Nutzungsproblemen indiziert (Hartson et al., 2001; Koutsabasis et al., 2007; Sears, 1997). Im Falle des Nutzertests mit retrospektivem Think-Aloud-Protokoll

wurde, wie bereits erwähnt, eine derartige Unterscheidung nicht vorgenommen und immer alle identifizierten Nutzungsprobleme zur Bestimmung der Gründlichkeit verwendet.

Explorative Datenanalyse zur post-hoc Erklärung von unerwarteten Ergebnissen Um unerwartete Ergebnisse bei der Bestimmung der Gründlichkeit und Gültigkeit (d.h. insbesondere Unterschiede zwischen der strikten und weniger strikten Anwendung von IntuiBeat-F) ausführlich post-hoc diskutieren zu können, wurden weitere nicht für die Hypothesen relevante Metriken berechnet. Einige Metriken widmeten sich dabei speziell IntuiBeat-F und dessen strikter Anwendung, wohingegen andere für beide Evaluationsmethoden berechnet wurden, da eine Berechnung ohne explizite Berücksichtigung von Rhythmus-Peaks bei IntuiBeat-F (weniger strikte Anwendung von IntuiBeat-F) für beide Methoden im Sinne einer explorativen Datenanalyse sinnvoll war.

- IntuiBeat-F: Anteil der nicht berücksichtigten Rhythmus-Peaks bei der Ableitung von allen Nutzungsproblemen an der Gesamtanzahl der von IntuiBeat-F generierten Rhythmus-Peaks (d.h. „In welchem Ausmaß blieben Rhythmus-Peaks bei der Ableitung von Nutzungsproblemen unberücksichtigt?“).
- IntuiBeat-F: Anteil der nicht berücksichtigten Rhythmus-Peaks bei der Ableitung von realen Nutzungsproblemen an der Gesamtanzahl der von IntuiBeat-F generierten Rhythmus-Peaks (d.h. „In welchem Ausmaß blieben Rhythmus-Peaks bei der Ableitung von realen Nutzungsproblemen unberücksichtigt?“).
- IntuiBeat-F: Anteil mithilfe von Rhythmus-Peaks identifizierter, kritischer Ereignisse an allen, auf alle Nutzungsprobleme hinweisende Ereignisse (d.h. „In welchem Ausmaß konnten alle kritischen Ereignisse, die auf ein Nutzungsproblem hinweisen, mithilfe von Rhythmus-Peaks identifiziert werden?“).
- IntuiBeat-F: Anteil mithilfe von Rhythmus-Peaks identifizierter, kritischer Ereignisse an allen, auf reale Nutzungsprobleme hinweisende Ereignisse (d.h. „In welchem Ausmaß konnten alle kritischen Ereignisse, die auf ein reales Nutzungsproblem hinweisen, mithilfe von Rhythmus-Peaks identifiziert werden?“).
- IntuiBeat-F: Anteil realer Nutzungsprobleme an allen Nutzungsproblemen, wenn für die Ableitung nur mithilfe von Rhythmus-Peaks identifizierte kritische Ereignisse genutzt werden (d.h. „In welchem Ausmaß halfen mithilfe von Rhythmus-Peaks identifizierte kritische Ereignisse bei der Ableitung von realen Nutzungsproblemen?“).
- IntuiBeat-F: Anzahl aller gefundenen Nutzungsprobleme, wenn für die Ableitung nur mithilfe von Rhythmus-Peaks identifizierte kritische Ereignisse genutzt werden (d.h. „Wie viele Nutzungsprobleme konnten mithilfe von Rhythmus-Peaks gefunden werden?“).
- IntuiBeat-F: Anzahl realer gefundener Nutzungsprobleme, wenn für die Ableitung nur mithilfe von Rhythmus-Peaks identifizierte kritische Ereignisse genutzt werden (d.h. „Wie viele reale Nutzungsprobleme konnte mithilfe von Rhythmus-Peaks gefunden werden?“).
- IntuiBeat-F: Trefferrate bei allen Nutzungsproblemen (d.h. Anteil von mithilfe von Rhythmus-Peaks identifizierten kritischen Ereignissen an allen dokumentierten Ereignissen, die zu einem Nutzungsproblem geführt haben).

- IntuiBeat-F: Trefferrate bei realen Nutzungsproblemen (d.h. Anteil von mithilfe von Rhythmus-Peaks identifizierten kritischen Ereignissen an allen dokumentierten Ereignissen, die zu einem realen Nutzungsproblem geführt haben).
- IntuiBeat-F: Fehlalarmrate bei allen Nutzungsproblemen (d.h. Anteil von mithilfe von Rhythmus-Peaks identifizierten kritischen Ereignissen an allen dokumentierten Ereignissen, die zu keinem Nutzungsproblem geführt haben).
- IntuiBeat-F: Fehlalarmrate bei realen Nutzungsproblemen (d.h. Anteil von mithilfe von Rhythmus-Peaks identifizierten kritischen Ereignissen an allen dokumentierten Ereignissen, die zu keinem realen Nutzungsproblem geführt haben).
- Beide Evaluationsmethoden: Anteil realer Nutzungsprobleme an allen Nutzungsproblemen, wenn für die Ableitung alle identifizierten kritischen Ereignisse genutzt werden (d.h. „In welchem Ausmaß halfen identifizierte kritische Ereignisse bei der Ableitung von realen Nutzungsproblemen?“).
- Beide Evaluationsmethoden: Anzahl aller gefundener Nutzungsprobleme, wenn für die Ableitung alle identifizierten kritischen Ereignisse genutzt werden (d.h. „Wie viele Nutzungsprobleme konnten gefunden werden?“).
- Beide Evaluationsmethoden: Anzahl realer gefundener Nutzungsprobleme, wenn für die Ableitung alle identifizierten kritischen Ereignisse genutzt werden (d.h. „Wie viele reale Nutzungsprobleme konnten gefunden werden?“).
- Beide Evaluationsmethoden: Trefferrate bei allen Nutzungsproblemen (d.h. Anteil von kritischen Ereignissen an allen dokumentierten Ereignissen, die zu einem Nutzungsproblem geführt haben).
- Beide Evaluationsmethoden: Trefferrate bei realen Nutzungsproblemen (d.h. Anteil von kritischen Ereignissen an allen dokumentierten Ereignissen, die zu einem realen Nutzungsproblem geführt haben).
- Beide Evaluationsmethoden: Fehlalarmrate bei allen Nutzungsproblemen (d.h. Anteil von kritischen Ereignissen an allen dokumentierten Ereignissen, die zu keinem Nutzungsproblem geführt haben).
- Beide Evaluationsmethoden: Fehlalarmrate bei realen Nutzungsproblemen (d.h. Anteil von kritischen Ereignissen an allen dokumentierten Ereignissen, die zu keinem realen Nutzungsproblem geführt haben).

Bewertung von zeitlicher Anwendungseffizienz

Als Operationalisierung der zeitlichen Anwendungseffizienz wurde im Rahmen des vierten Experiments in Anlehnung an Akers (2010), welcher eine ähnliche Operationalisierung für die Bewertung der zeitlichen Anwendungseffizienz für seine formative Methode nutzte, die durchschnittliche Zeit für die Identifikation eines echten Nutzungsproblems gewählt. Zur Berechnung dieses Durchschnittswerts wurde die gesamte Anwendungszeit in Sekunden, die die Summe aus der Instruktionszeit des Nutzertests (z.B. Instruktion und Baselineerhebung bei IntuiBeat-F), der gesamten Durchführungszeit der Systemnutzung und der Auswertungszeit des Nutzertests (d.h. Extraktion der Nutzungsprobleme durch die getesteten Methoden, siehe Abbildung 6.2) darstellte, durch die Anzahl der mit der jeweiligen formativen Evaluationsmethode gefundenen realen Nutzungsprobleme

geteilt (AV_ZeitlicheAnwendungseffizienz). Eine geringe durchschnittliche Identifikationszeit indiziert dabei eine hohe Anwendungseffizienz der jeweiligen formativen Evaluationsmethode, wohingegen eine hohe durchschnittliche Identifikationszeit eine geringe Anwendungseffizienz der jeweiligen formativen Evaluationsmethode signalisiert. Wie bereits bei der Operationalisierung der zeitlichen Anwendungseffizienz bei der Meta-Evaluation von IntuiBeat-S (siehe Teilabschnitt 5.2.3.3) wurden auch beim formativen Teil keine Zeit für die Administration von Fragebögen, Erhebung von demographischen Daten oder Ähnlichem (z.B. allgemeine Einführung des Experiments) berücksichtigt. Wie bereits beim zweiten Experiment wurde Zeit für die Aufgabenbearbeitung aus den Videoaufzeichnungen (d.h. Länge der Videos) entnommen. Der Beginn der Sitzung wurde aus dem Probandensystem entnommen, da die Versuchspersonen ja einzeln zu verschiedenen Zeitpunkten bestellt wurden. Entsprechende Abweichungen (z.B. Verspätungen), die Zeiten für allgemeine Einführungen, sowie Fragebogenerhebungen wurden hierbei vom Evaluator notiert und bei der Berechnung berücksichtigt und herausgerechnet. Der Beginn der Baseline-Erhebung zur Kalkulation der Dauer der Instruktion wurde aus der Erstellungszeit der entsprechenden Base-Datei abgeleitet. Die Zeit für das retrospektive Interview und damit für die Auswertung des Nutzertests wurde vom Evaluator händisch gestoppt.

An dieser Stelle sei anzumerken, dass hierbei nicht die Zeit für die Ableitung der Nutzungsprobleme aus den kritischen Ereignissen mithilfe des Verfahrens von Burmester (2016) und die Entscheidung über deren Echtheit mithilfe der handlungsorientierten Fehlertaxonomie (Zapf et al., 1989) berücksichtigt wurde, da diese Schritte in beiden getesteten formativen Evaluationsmethoden erfolgten.

Demographische Variablen

Als soziodemografische Daten wurden von den Versuchspersonen Alter, Geschlecht und Studiengang erhoben.

Vorerfahrung bei der Nutzung von CUIs

Die Vorerfahrung der Versuchspersonen bezüglich der industriellen Regalplanungssoftware *IPO.Rack* wurde mit einem hierfür konstruierten TFQ als Kontrollvariable erhoben (KV_Vorerfahrung), welcher, wie in den Experimenten zuvor, in Papierform administriert wurde. Als Items für diesen Fragebogen wurden lediglich „IPO.Rack“ selbst und „Andere Regalplanungssoftware?“ genutzt, wobei letztere eine Wildcard darstellte und die Angabe einer beliebigen bekannten Regalplanungssoftware erlaubte. Es wurde bei der Konstruktion des TFQ im Vergleich zu einigen TFQs aus den vorigen Experimenten auf die Nennung konkreter CUIs als Beispiele verzichtet, da Spezialsoftware für die Planung von Industrieregalsystemen mit sehr hoher Wahrscheinlichkeit für die als Stichprobe fungierenden Studierenden der Medienkommunikation und Mensch-Computer-Systeme unbekannt ist. Falls es doch zu einer Ausnahme kommt, könnte die Expertise der Versuchspersonen mit einer anderen Regalplanungssoftware mithilfe der Wildcard entsprechend erfasst werden.

Wie bei den anderen konstruierten TFQs, beurteilten die Versuchspersonen diese beiden Items dann bezüglich deren Nutzungshäufigkeit (Wertebereich: 0 - 6) und des genutzten

Funktionsumfangs (Wertebereich: 0 - 4). Im Anschluss wurde dann für jede Dimension (d.h. Nutzungshäufigkeit und Funktionsumfang) ein Mittelwert gebildet und daraufhin ein Gesamtmittelwert über beide Mittelwerte als Operationalisierung der *Vorerfahrung bei der Nutzung von CUIs* berechnet (Wertebereich: 0 - 5). Die Entscheidung für eine Mittelwertbildung anstelle einer Summenberechnung wurde im Rahmen des ersten Experiments ausführlich begründet, weswegen an dieser Stelle lediglich darauf verwiesen wird (siehe Absatz „Vorerfahrung bei der Nutzung von CUIs“ des Teilabschnitts 5.1.4.3). Da kein einheitlicher TFQ existiert und dieser jeweils für die getesteten Softwareanwendungen erstellt werden musste, ist der in diesem Experiment genutzte TFQ vollständig in Anhang B.4.2 dieser Arbeit zu finden.

Apparatur

Die im Rahmen des vierten Experiments verwendete Apparatur unterschied sich nur in einigen Punkten von der im ersten Experiment genutzten Apparatur (siehe Abbildung 6.4). Anstelle des MSI GT62VR Gaming Notebooks kam als Versuchspersonen-PC ein Fujitsu Esprimo P710 (Windows 7 Home, 3.4 GHz Dual-Core Intel i3, 4 GB Arbeitsspeicher, Intel HD Graphics 2500 Grafikkarte, Auflösung: 1920 x 1080 Pixel) zum Einsatz. Schließlich wurden für das vierte Experiment keine Kopfhörer benötigt, weil die rhythmische Wahrnehmung nicht mithilfe eines PROMS abgefragt wurde. Die restliche Apparatur war identisch zum ersten Experiment, weswegen für weitere Details auf den entsprechenden Absatz des ersten Experiments (siehe Absatz 5.1.4.3) verwiesen werden soll.

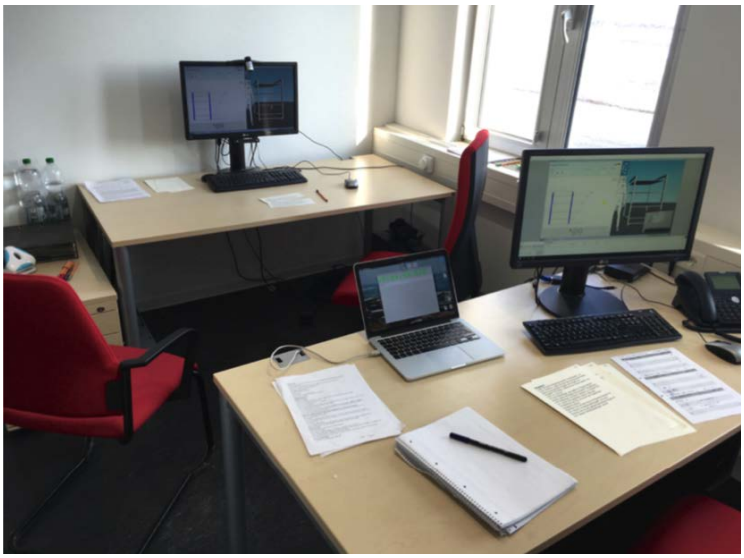


Abbildung 6.4. Apparatur des vierten Experiments, bestehend aus Versuchspersonen-PC (hinten), Analyse-PC (vorne links) und Evaluatoren-PC (vorne rechts).

6.1.3.4 Versuchsdurchführung

Wie im ersten und zweiten Experiment erfolgte auch beim vierten Experiment die Rekrutierung der Versuchspersonen über das Probanden-System des Instituts für Mensch-

Computer-Medien der Universität Würzburg. Das Experiment wurde darin als „Nutzertest verschiedener Software-Anwendungen“ beworben und wie im ersten Experiment beschrieben (siehe Teilabschnitt 5.1). Alle angebotenen Termine waren ebenfalls Einzeltermine, da auch immer nur eine Person gleichzeitig getestet wurde.

Das vierte Experiment wurden im selben kleinen, reizabgeschirmten, stimulusarmen Labor des Lehrstuhls für Psychologische Ergonomie der Universität Würzburg durchgeführt, das bereits für das erste Experiment genutzt wurde (siehe Abbildung 5.7). Der Geräuschpegel von außen war über den gesamten Erhebungszeitraum konstant niedrig. Bevor die Versuchspersonen das Testlabor betreten durften, wurden sie vom Evaluator mündlich aufgefordert ihre Mobiltelefone in den Flugmodus zu schalten, um eine Ablenkung durch diese während der Untersuchung zu vermeiden. Die eigentliche Datenerhebung der Studie unterteilte sich in drei Phasen (d.h. Vorphase, Hauptphase, Abschlussphase), die insgesamt circa 60 Minuten in Anspruch nahmen.

Im Folgenden soll nur auf die Unterschiede in den einzelnen Phasen der Versuchsdurchführung im Vergleich zum ersten Experiment eingegangen werden und für weitere Details bezüglich der Gemeinsamkeiten auf die Erläuterung im Rahmen des ersten Experiments in Teilabschnitt 5.1 verwiesen werden. An dieser Stelle sei weiter darauf hingewiesen, dass zusätzlich eine SEA-Skala nach jeder erledigten Aufgabe erhoben wurde und das Ausmaß der intuitiven Systemnutzung bei jeder Software abschließend mithilfe eines QUESI beurteilt wurde. Jedoch wurden diese Daten nicht im Zuge der Meta-Evaluation von IntuiBeat-F genutzt, weswegen auch nicht weiter darauf eingegangen wird, sondern nur aus Gründen der Replizierbarkeit hiervon an dieser Stelle berichtet wird.

Vorphase (5 Minuten)

Wie bereits bei den vorherigen beiden Experimenten, wurde im Rahmen der Vorphase des vierten Experiments nicht mehr die rhythmische Wahrnehmung der Versuchspersonen erhoben, weswegen sich die Vorphase des vierten Experiments identisch zum zweiten Experiment gestaltete. Die Vorerfahrung der Versuchspersonen bei der Nutzung von CUIs wurde lediglich mit dem auf die Regalplanungssoftware angepassten TFQ erhoben, weswegen für genauere Details bezüglich der Vorphase auf den Teilabschnitt 5.2.3.4 des zweiten Experiments verwiesen wird.

Hauptphase (50 Minuten)

Um die wissenschaftliche Güte von IntuiBeat-F als formative Evaluationsmethode für intuitive Benutzung untersuchen zu können, wurden die Versuchspersonen zu Beginn der Hauptphase einer der beiden Versuchsbedingungen, *IntuiBeat-F* und *Nutzertest mit retrospektivem Think-Aloud-Protokoll*, randomisiert zugewiesen und diese Zuweisung über alle Versuchspersonen hinweg ausbalanciert.

Versuchspersonen in der Versuchsbedingung *Nutzertest mit retrospektivem Think-Aloud-Protokoll* erhielten zunächst eine standardisierte mündliche Instruktion durch den Evaluator für die Hauptphase der Untersuchung. Im Rahmen dieser Instruktion wurde den

Versuchspersonen zunächst der genaue Ablauf des Nutzertests mit retrospektivem Think-Aloud-Protokoll mitgeteilt und ihnen erläutert, dass das Ziel eines solchen Tests darin bestehe, sowohl Probleme des Systems zu identifizieren als auch das Ausmaß an intuitiver Benutzung der CUIs zu bewerten. Daraufhin wurde den Versuchspersonen die Regalplanungssoftware *IPO.Rack* vom Evaluator standardisiert mündlich vorgestellt und das fiktive Anlagenplanungsszenario präsentiert. Anschließend händigte der Evaluator den Versuchspersonen die Instruktion der ersten Testaufgabe und die Beschreibung des Szenarios auf Papier aus. Daraufhin startete er die Software *IPO.Rack*.

Nachdem die Versuchspersonen dem Evaluator durch ein Handzeichen signalisiert hatten, dass sie mit der Bearbeitung der ersten Aufgabe beginnen möchten, wies der Evaluator auf die stillschweigende Bearbeitung der Aufgabe hin und startete daraufhin die Bildschirmaufzeichnung. Wie bereits erwähnt, fand während der eigentlichen Aufgabenbearbeitung in dieser Bedingung eine Protokollierung kritischer Ereignisse durch den Evaluator statt. Nach der Erledigung der ersten Aufgabe gaben die Versuchspersonen dem Evaluator erneut ein Handzeichen und dieser stoppte daraufhin die Bildschirmaufzeichnung. Da alle Testaufgaben so ausgewählt wurden, dass ein Expertennutzer eine Testaufgabe in einer Minute erledigen kann, wurde die Aufgabenbearbeitung vom Versuchsleiter nach fünf Minuten abgebrochen. Nachdem die Versuchspersonen die erste Aufgabe abgeschlossen hatten, händigte der Evaluator ihnen die Instruktion der zweiten Aufgabe auf Papier aus und bat sie die Instruktion aufmerksam zu lesen. Die geschilderte Bearbeitungsprozedur wurde für diese und die restlichen Testaufgaben wiederholt. Abschließend erfolgte noch das retrospektive Interview, um Nutzungsprobleme bei der Systemnutzung anhand der vorgespielten Bildschirmaufzeichnung identifizieren zu können. Nachdem die Versuchspersonen die Bearbeitung aller Aufgaben mit IPO.Rack abgeschlossen hatten, endete die Hauptphase der Versuchsbedingung *Nutzertest mit retrospektivem Think-Aloud-Protokoll*. Die Hauptphase des Experiments unter Berücksichtigung dieser Bedingung erstreckte sich auf circa 50 Minuten, wobei circa 30 Minuten für die retrospektive Befragung aufgewendet wurden.

Versuchspersonen in der Versuchsbedingung *IntuiBeat-F* erhielten eine ähnliche standardisierte mündliche Instruktion durch den Evaluator wie Versuchspersonen, die der Versuchsbedingung *Nutzertest mit retrospektivem Think-Aloud-Protokoll* zugeordnet wurden. Zusätzlich zu der üblichen Instruktion über den genauen Ablauf des Nutzertests, wurde ihnen vom Evaluator, wie beim ersten Experiment (siehe Teilabschnitt 5.1), zusätzlich mitgeteilt, dass die intuitive Benutzung mithilfe einer Rhythmuszweitaufgabe im Rahmen des Experiments gemessen werden soll. Dementsprechend wurde von Versuchspersonen im Anschluss eine Rhythmus-Baseline erhoben, wie es im Detail bereits beim ersten Experiment (siehe Teilabschnitt 5.1) beschrieben wurde. Wie schon im Rahmen der Beschreibung der Versuchsdurchführung des Nutzertests mit retrospektivem Think-Aloud-Protokoll erläutert, bearbeiteten die Versuchspersonen auch in der Versuchsbedingung *IntuiBeat-F* die experimentellen Aufgaben, wobei sie zusätzlich zuvor instruiert wurden, den gelernten Rhythmus mit einem USB-Fußpedal während der Systemnutzung zu klopfen. Der Evaluator musste deswegen, wie bereits beim ersten Experiment (siehe Teilabschnitt 5.1), neben der Bildschirmaufzeichnung auch gleichzeitig die Rhythmusaufzeichnung zu Beginn einer jeden Testaufgabe starten und zum Ende jeder Testaufgabe beenden.

Wie bereits erwähnt, fand in dieser Versuchsbedingung während der eigentlichen Aufgabenbearbeitung keine Protokollierung kritischer Ereignisse durch den Evaluator statt. Falls der Evaluator dennoch einige Ereignisse notierte, wurden diese bei der späteren Datenauswertung im Zuge der Meta-Evaluation von IntuiBeat-F nicht berücksichtigt und lediglich für das Projekt 3D-GUIde verwendet. Nachdem die Versuchspersonen auch in dieser Ausführung alle Aufgaben mit IPO.Rack bearbeitet hatten, endete die Hauptphase der Versuchsbedingung *IntuiBeat-F*. Die Hauptphase des Experiments unter Berücksichtigung dieser Bedingung war circa fünf Minuten länger, da diese Zeit im Vergleich zum Nutzertest mit retrospektivem Think-Aloud-Protokoll für das Erheben der Rhythmus-Baseline aufgewendet wurde. Die retrospektive Befragung dauerte wie in der anderen Bedingung circa 30 Minuten.

Abschlussphase (5 Minuten)

Die Abschlussphase des vierten Experiments war mit der Abschlussphase des zweiten Experiments identisch, weswegen bezüglich der Gemeinsamkeiten bei der Gestaltung der Abschlussphase für genauere Details auf den Teilabschnitt 5.2.3.4 des zweiten Experiments verwiesen wird.

6.1.3.5 Statistische Auswertung

Die statistische Datenauswertung erfolgte mittels IBM SPSS Statistics 25 für macOS. Zunächst wurde vor der eigentlichen Datenauswertung geprüft, ob die Daten Ausreißer enthielten, die Datensätze der erhobenen Stichprobe vollständig und die Voraussetzungen der statistischen Tests erfüllt waren. Alle statistischen Tests und Analysen der Teststärke erfolgten zweiseitig.

Überprüfung der Gründlichkeit von IntuiBeat-F

Die Überprüfung der Gründlichkeit (Hypothese H1) erfolgte univariat in einem einfaktoriellen Between-Subjects Design durch einen t-Test für unabhängige Stichproben. Als unabhängige Variable fungierte die *Art der formativen Evaluationsmethode* mit den Ausprägungen: IntuiBeat-F vs. Nutzertest mit retrospektivem Think-Aloud-Protokoll. Als abhängige Variable kam der *Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der vorhandenen realen Nutzungsprobleme* (d.h. *Gründlichkeit*) zum Einsatz, wobei sich dieser bei IntuiBeat-F einmal weniger strikt auf Basis aller Nutzungsprobleme (H1.A) und einmal strikt auf Basis ausschließlich algorithmisch mithilfe von Rhythmus-Peaks identifizierter Nutzungsprobleme (H1.B) berechnete.

Da die *Vorerfahrung bei der Nutzung von CUIs*, die als Kontrollvariable erhoben wurde, einen ungewollten Einfluss auf die Gründlichkeit der verglichenen Methoden haben könnte, musste eine derartige Konfundierung vor der Überprüfung der Gründlichkeit von IntuiBeat-F ausgeschlossen werden. Hierzu wurde ein t-Test für unabhängige Stichproben bezüglich der Vorerfahrung, sowie Pearson-Produkt-Moment-Korrelationen zwischen

der Vorerfahrung und der Gründlichkeit innerhalb beider Ausprägungen der unabhängigen Variable gerechnet (siehe Teilabschnitt 6.1.4.2). Da hierbei keine signifikanten Ergebnisse festgestellt werden konnten (d.h. kein ungewollter Einfluss der Vorerfahrung auf die Gründlichkeit), wurde von einer univariaten Kovarianzanalyse (ANCOVA) abgesehen (siehe Döring & Bortz, 2016; Field, 2017) und sich stattdessen für eine Datenauswertung mithilfe des erwähnten t-Tests für unabhängige Stichproben ohne Berücksichtigung der Vorerfahrung bei der Nutzung von CUIs als Kovariate entschieden (siehe Teilabschnitt 6.1.4.3). Das Signifikanzniveau für alle damit verbundenen statistischen Berechnungen betrug $\alpha = .05$ und die berechneten Effektgrößen wurden als Absolutwerte berichtet. Die Effektstärke für die t-Tests wurde mithilfe von G*Power (Faul et al., 2009) berechnet. Dabei indizierten absolute Effektstärken d um .20 einen kleinen, um .50 einen mittleren und um .80 einen großen Effekt (J. Cohen, 1992). Ein absoluter Korrelationskoeffizient $|r|$ um .10 wurde als ein schwacher, ein Korrelationskoeffizient $|r|$ um .30 als ein moderater, und ein Korrelationskoeffizient $|r|$ um .50 als hoher Zusammenhang interpretiert (J. Cohen, 1988). Die Überprüfung der statistischen Voraussetzungen der Pearson-Produkt-Moment-Korrelationen und der t-Tests werden im folgenden Ergebnisteil berichtet (siehe Teilabschnitt 6.1.4.1).

Überprüfung der Gültigkeit von IntuiBeat-F

Die Überprüfung der Gültigkeit (Hypothese H2) erfolgte ebenfalls univariat in einem einfaktoriellen Between-Subjects Design durch einen t-Test für unabhängige Stichproben. Als unabhängige Variable fungierte die *Art der formativen Evaluationsmethode* mit den Ausprägungen: IntuiBeat-F vs. Nutzertest mit retrospektivem Think-Aloud-Protokoll. Als abhängige Variable kam der *Anteil von gefundenen realen Nutzungsproblemen an der Gesamtzahl der von der jeweiligen Evaluationsmethode gefundenen Nutzungsprobleme* (d.h. *Gültigkeit*) zum Einsatz, wobei sich dieser bei IntuiBeat-F einmal weniger strikt auf Basis aller Nutzungsprobleme (H2.A) und einmal strikt auf Basis ausschließlich algorithmisch mithilfe von Rhythmus-Peaks identifizierter Nutzungsprobleme (H2.B) berechnete.

Wie bereits bei der Überprüfung der Gründlichkeit beschrieben, kann die als Kontrollvariable erfasste *Vorerfahrung bei der Nutzung von CUIs* einen ungewollten Einfluss auf die Gründlichkeit nehmen, was auch bei der Überprüfung der Gültigkeit aufgrund nicht signifikanter Ergebnisse (siehe Teilabschnitt 6.1.4.2) ausgeschlossen werden konnte (d.h. kein ungewollter Einfluss der Vorerfahrung auf die Gültigkeit). Demzufolge kam lediglich ein t-Test für unabhängige Stichproben anstelle einer univariaten Kovarianzanalyse (ANCOVA) für die Überprüfung der Gültigkeit zum Einsatz (siehe Teilabschnitt 6.1.4.4). Die statistische Auswertung bei der Überprüfung der Gültigkeit war ansonsten identisch mit dem Vorgehen bei der Überprüfung der Gründlichkeit, weswegen für Details, wie die Interpretation von Effektgrößen, auf den entsprechenden Absatz verwiesen wird. Die Überprüfung der statistischen Voraussetzungen der Pearson-Produkt-Moment-Korrelationen und der t-Tests werden im folgenden Ergebnisteil berichtet (siehe Teilabschnitt 6.1.4.1).

Überprüfung der zeitlichen Anwendungseffizienz von IntuiBeat-F

Die Überprüfung der zeitlichen Anwendungseffizienz (Hypothese H3) von IntuiBeat-F erfolgte univariat in einem einfaktoriellen Between-Subjects Design durch einen t-Test für unabhängige Stichproben. Als unabhängige Variable fungierte die *Art der formativen Evaluationsmethode* mit den Ausprägungen: IntuiBeat-F vs. Nutzertest mit retrospektivem Think-Aloud-Protokoll. Als abhängige Variable kam die *zeitliche Anwendungseffizienz* zum Einsatz.

Da aufgrund der Operationalisierung der zeitlichen Anwendungseffizienz als durchschnittliche Zeit zur Identifikation eines mithilfe der jeweiligen Methode gefundenen realen Nutzungsproblems, auch die Anzahl der durch die jeweilige Methode gefundenen realen Nutzungsprobleme bestimmt werden musste, die wie bereits bei der Überprüfung der Gründlichkeit und Gültigkeit ungewollt von der *Vorerfahrung bei der Nutzung von CUIs* beeinflusst werden konnte, musste eine solche Konfundierung im Vorfeld ebenfalls ausgeschlossen werden. Hierzu wurde wie zuvor ein t-Test für unabhängige Stichproben bezüglich der Kontrollvariablen, sowie Pearson-Produkt-Moment-Korrelationen zwischen der Vorerfahrung und der zeitlichen Anwendungseffizienz innerhalb beider Ausprägungen der unabhängigen Variable gerechnet (siehe Teilabschnitt 6.1.4.2). Da hierbei keine signifikanten Ergebnisse festgestellt werden konnten (d.h. kein ungewollter Einfluss der Vorerfahrung auf die zeitliche Anwendungseffizienz), wurde von einer univariaten Kovarianzanalyse (ANCOVA) abgesehen (Döring & Bortz, 2016; Field, 2017) und sich stattdessen für eine Datenauswertung mithilfe des erwähnten t-Tests für unabhängige Stichproben ohne Berücksichtigung der Vorerfahrung bei der Nutzung von CUIs als Kovariate entschieden (siehe Teilabschnitt 6.1.4.5). Die statistische Auswertung bei der Überprüfung der zeitlichen Anwendungseffizienz war ansonsten identisch mit dem Vorgehen bei der Überprüfung der Gründlichkeit, weswegen für Details, wie die Interpretation von Effektgrößen, auf den entsprechenden Absatz verwiesen wird. Die Überprüfung der statistischen Voraussetzungen der Pearson-Produkt-Moment-Korrelationen und der t-Tests werden im folgenden Ergebnisteil berichtet (siehe Teilabschnitt 6.1.4.1).

6.1.4 Ergebnisse

Im folgenden Abschnitt werden die Ergebnisse bezüglich der in den Teilabschnitten 6.1.1 und 6.1.2 beschriebenen Hypothesen deskriptiv und inferenzstatistisch berichtet. Vor der eigentlichen Datenanalyse wird zunächst auf die Überprüfung der statistischen Voraussetzungen eingegangen. Dabei wurde bei allen abhängigen Variablen und Kontrollvariablen ein metrisches Skalenniveau angenommen.

6.1.4.1 Überprüfung der statistischen Voraussetzungen

Überprüfung von Ausreißern

Zur Überprüfung univariater Ausreißer wurden vor der Untersuchung der einzelnen Hypothesen modifizierte z-Werte herangezogen, da die Verwendung von z-Werten generell

ein häufig angewendetes Verfahren zur Identifikation univariater Ausreißer darstellt (Cousineau & Chartier, 2010; Shiffler, 1988). Es wurde sich speziell für die Verwendung von modifizierten z-Werten entschieden, da diese auch bei Stichproben mit geringer Größe mit hoher Zuverlässigkeit funktionieren (Garcia, 2012; Iglewicz & Hoaglin, 1993; Seo, 2006). Als univariate Ausreißer wurden in Anlehnung an Iglewicz und Hoaglin (1993) Werte identifiziert, deren absoluter modifizierter z-Wert größer als 3.5 lag. Es mussten auf diese Weise bei keiner der analysierten abhängigen Variablen und Kontrollvariablen Werte ausgeschlossen werden.

An dieser Stelle ist jedoch anzumerken, dass aufgrund der geringen Varianz von KV_Vorerfahrung ($M = .07$, $SD = .25$) und der damit verbundenen Tatsache, dass alle Versuchspersonen nahezu keine Vorerfahrung bei der Nutzung von CUIs aufwiesen und die angegebenen Werte damit fast das Minimum des TFQ (d.h. Minimum des TFQ lag bei 0) repräsentierten, in diesem Fall kein modifizierter z-Wert berechnet werden konnte (d.h. Dividende durch Null). Stattdessen wurde anhand eines Boxplots von KV_Vorerfahrung überprüft, inwiefern darin univariate Ausreißer zu finden sind (siehe Howell, 2009). Mit diesem Verfahren konnte zwei Ausreißer ($\text{Ausreißer}_1 = .75$, $\text{Ausreißer}_2 = 1$) festgestellt werden, welche aber aufgrund der geringen Stichprobengröße und der Tatsache, dass ansonsten die Vorerfahrung konstant bei Null gewesen wäre, nicht ausgeschlossen wurden.

Überprüfung der Voraussetzung der Normalverteilung

Die univariate Normalverteilung der abhängigen Variablen und Kontrollvariablen wurde für jede Ausprägung der unabhängigen Variable mittels Kolmogorov-Smirnov-Tests ($p \geq .05$, siehe Field, 2017) und Sichtprüfung anhand eines Q-Q-Diagramms geprüft. Dabei konnten auf Basis dieser beiden Kriterien bei AV_Gültigkeit_Strickt im Rahmen der Überprüfung der Gültigkeit (Hypothese H2) und bei KV_Vorerfahrung keine Normalverteilung in beiden Gruppen festgestellt werden. Aufgrund der Tatsache, dass ein ungepaarter t-Test bei etwa gleich großen Gruppen robust gegenüber Verletzungen der Normalverteilungsannahme ist (Glass et al., 1972; Pagano, 2006; Salkind, 2010; Wilcox, 2011), wurde sich für eine eindeutige Interpretation und Vergleichbarkeit der Ergebnisse (d.h. Mittelwerte statt Medianen) der verschiedenen Experimente gegen Transformationen und entsprechende nonparametrische Verfahren zur Überprüfung der Gültigkeit (Hypothese H2) entschieden.

Überprüfung der Voraussetzung der Homoskedastizität

Zur Überprüfung der Homoskedastizität zwischen den Ausprägungen der unabhängigen Variable kamen Levene-Tests zum Einsatz (siehe Field, 2017), welche im Rahmen der Überprüfung der Gründlichkeit (Hypothese H1), der Überprüfung der Gültigkeit (Hypothese H2) und der Überprüfung der zeitlichen Anwendungseffizienz (Hypothese H3) gerechnet und bei allen abhängigen Variablen und Kontrollvariablen ($p \geq .05$, siehe Field, 2017) Varianzhomogenität bestätigen konnten.

6.1.4.2 Überprüfung einer möglichen Konfundierung durch die Vorerfahrung bei der Nutzung von CUIs

Die Vorerfahrung bei der Nutzung von CUIs (Wertebereich: 0 - 5) unterschied sich nicht signifikant zwischen den beiden Ausprägungen der unabhängigen Variable (Nutzertest mit retrospektivem Think-Aloud-Protokoll: $M = .06$; $SD = .22$; IntuiBeat-F: $M = .08$, $SD = .29$), $t(22) = .200$, $p = .843$, d = .08. Laut J. Cohen (1988) kann dieser Effekt als klein interpretiert werden ($d \leq .5$). Eine konservative post-hoc Analyse der Teststärke bezüglich KV_Vorerfahrung mithilfe von G*Power (Faul et al., 2009) mit $df = 22$ und einer angenommenen geringen Effektstärke (d = .08) ergab lediglich eine geringe Teststärke (1 - $\beta = .05$) im Zuge der Auswertung der Vorerfahrung bei der Nutzung von CUIs. Um jedoch eine ausreichend große Power (1 - $\beta \geq .80$) beim festgestellten Effekt erzielen zu können, wären pro Gruppe 2601 Versuchspersonen nötig gewesen, was aufgrund des straffen Zeitplans im Anwenderprojekt, des dortigen Fokus auf qualitative Ergebnisse, des Verständnisses der Meta-Evaluation von IntuiBeat-F als Nebenprodukt und der personellen Einschränkungen nicht möglich gewesen wäre. Auch bei Annahme einer großen Effektstärke (d = .8) hätten immerhin noch 26 Versuchspersonen pro Gruppe getestet werden müssen, was im Hinblick auf die Einschränkungen durch das Anwenderprojekt nicht notwendig erschien.

Da die erhobene Vorerfahrung bei der Nutzung von CUIs dennoch einen ungewollten Einfluss auf den Vergleich der beiden Ausprägungen der unabhängigen Variable haben könnte, wurde für die Kontrollvariable mithilfe von Pearson-Produkt-Moment-Korrelationen sichergestellt, dass kein Zusammenhang zwischen dieser und den abhängigen Variablen innerhalb der beiden Ausprägungen der unabhängigen Variable besteht ($p > .05$). Die Kontrollvariable KV_Vorerfahrung wurde dementsprechend nicht als Kovariate in der statistischen Auswertung berücksichtigt, da kein linearer Zusammenhang zwischen dieser und den abhängigen Variablen bestand (Döring & Bortz, 2016; Field, 2017).

6.1.4.3 Überprüfung der Gründlichkeit

Wie erwartet, lag die Gründlichkeit (%) von IntuiBeat-F ($M = .32$, $SD = .08$) signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll ($M = .22$, $SD = .07$), wenn alle mit IntuiBeat-F abgeleiteten Nutzungsprobleme berücksichtigt (d.h. weniger strikte Anwendung von IntuiBeat-F) wurden (H1.A), $t(22) = -3.28$, $p = .003$, d = 1.34. Es war laut J. Cohen (1988) ein großer Effekt feststellbar (d $\geq$ .8). Entgegen der Erwartungen, konnte jedoch kein signifikanter Unterschied zwischen IntuiBeat-F ($M = .19$, $SD = .11$) und dem Nutzertest mit retrospektivem Think-Aloud-Protokoll ($M = .23$, $SD = .07$) festgestellt werden, wenn nur die auf Basis des Algorithmus von IntuiBeat-F abgeleiteten Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F) berücksichtigt wurden (H1.B), $t(22) = 1.20$, $p = .244$, d = .43. Es war laut J. Cohen (1988) ein kleiner Effekt feststellbar (d $\leq$ .5).

Eine konservative post-hoc Analyse der Teststärke mithilfe von G*Power (Faul et al., 2009) bezüglich der AV_Gründlichkeit_Strikt mit $df = 22$ und einer angenommenen kleinen Effektstärke (d = .43) ergab eine geringe Teststärke (1 - $\beta = .17$). Um jedoch eine ausreichend große Power (1 - $\beta \geq .80$) beim festgestellten Effekt erzielen zu können, wären pro Gruppe 85 Versuchspersonen nötig gewesen, was aufgrund des straffen Zeitplans im

Anwenderprojekt, des dortigen Fokus auf qualitative Ergebnisse, des Verständnisses der Meta-Evaluation von IntuiBeat-F als Nebenprodukt und der personellen Einschränkungen im Projekt 3D-GUIde nicht möglich gewesen wäre. Auch bei Annahme einer großen Effektstärke (d = .8) hätten immerhin noch 26 Versuchspersonen pro Gruppe getestet werden müssen, was im Hinblick auf die Einschränkungen durch das Anwenderprojekt nicht notwendig erschien.

6.1.4.4 Überprüfung der Gültigkeit

Wie erwartet, lag die Gültigkeit (%) von IntuiBeat-F (M = .92, SD = .08) signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll (M = .81, SD = .13), wenn alle mit IntuiBeat-F abgeleiteten Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) berücksichtigt wurden (H2.A), $t(22)$ = - 2.66, p = .014, d = 1.09. Es war laut J. Cohen (1988) ein großer Effekt feststellbar (d $\geq$ .8). Auch wenn ausschließlich die auf Basis des Algorithmus von IntuiBeat-F abgeleiteten Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F) berücksichtigt wurden (H2.B), lag die Gültigkeit, wie erwartet, von IntuiBeat-F (M = .96, SD = .10) signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll (M = .81, SD = .13), $t(22)$ = - 3.26, p = .004, d = 1.33. Es war laut J. Cohen (1988) hier ebenfalls ein großer Effekt feststellbar (d $\geq$ .8).

6.1.4.5 Überprüfung der zeitlichen Anwendungseffizienz

Wie erwartet, lag die zeitliche Anwendungseffizienz (s) von IntuiBeat-F in Form einer geringeren durchschnittlichen Problemidentifikationszeit für reale Nutzungsprobleme (M = 193.84, SD = 56.35) signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll (M = 247.51, SD = 53.48) , $t(22)$ = 2.39, p = .026, d = .98. Es war laut J. Cohen (1988) ein großer Effekt feststellbar (d $\geq$ .8).

6.1.4.6 Explorative Datenanalyse

Wie bereits in Teilabschnitt 6.1.3.3 dieses Experiments angesprochen, wurden zusätzlich eine Reihe von Metriken für eine explorative Datenanalyse berechnet, mit deren Hilfe verstanden werden kann, warum es zu den Unterschieden bezüglich Gründlichkeit (siehe Teilabschnitt 6.1.4.3) und Gültigkeit (siehe Teilabschnitt 6.1.4.4) bei der strikten und weniger strikten Anwendung von IntuiBeat-F gekommen ist. Im Folgenden sollen diese Metriken nun berichtet werden.

Tabelle 6.2. *Metriken zur explorativen Datenanalyse von IntuiBeat-F und des Nutzertests mit retrospektivem Think-Aloud-Protokoll im Rahmen des vierten Experiments (NPs: Nutzungsprobleme; KEs: kritische Ereignisse; RP-KEs: mit Rhythmus-Peaks assoziierte kritische Ereignisse; Alle NPs aus KEs ≈ weniger strikte Anwendung von IntuiBeat-F; Nur NPs aus RP-KEs ≈ strikte Anwendung von IntuiBeat-F).*

Metrik	Experiment 4	
	IntuiBeat-F	Nutzertest
Trefferrate NPs (%):		
- Alle NPs aus KEs	93.09	88.26
- Nur NPs aus RP-KEs	100	-
Trefferrate reale NPs (%):		
- Alle NPs aus KEs	88.16	87.05
- Nur NPs aus RP-KEs	95.75	-
Fehlalarmrate NPs (%):		
- Alle NPs aus KEs	6.91	11.74
- Nur NPs aus RP-KEs	0	-
Fehlalarmrate reale NPs (%):		
- Alle NPs aus KEs	11.84	12.95
- Nur NPs aus RP-KEs	4.25	-
Anzahl NPs:		
- Alle NPs aus KEs	60	54
- Nur NPs aus RP-KEs	41	-
Anzahl realer NPs:		
- Alle NPs aus KEs	56	51
- Nur NPs aus RP-KEs	40	-
Anteil realer NPs an NPs (%):		
- Alle NPs aus KEs	93.33	94.44
- Nur NPs aus RP-KEs	97.56	-

Explorative Datenanalyse von IntuiBeat-F (weniger strikte Anwendung)

Mithilfe von IntuiBeat-F konnten über alle Versuchspersonen hinweg 304 kritische Ereignisse erkannt werden, wovon unter Berücksichtigung der, in der Abbildung 6.2 in Teilabschnitt 6.1.3 beschriebenen, Arbeitsschritte zur Ableitung von Nutzungsproblemen (Burmester, 2016) und der handlungsorientierten Fehlertaxonomie (Zapf et al., 1989) 283 kritische Ereignisse (93.09 %) zu insgesamt 60 einzigartigen Nutzungsproblemen konsolidiert

wurden (d.h. weniger strikte Anwendung von IntuiBeat-F). Von diesen 60 Nutzungsproblemen wurde bei vieren, die sich aus 15 von den 283 kritischen Ereignissen (5.30 %) ableiteten, ein Bewegungsfehler als Ursache klassifiziert. Es handelt sich dementsprechend nur bei 56 der insgesamt abgeleiteten 60 Nutzungsprobleme (93.33 %) um reale Nutzungsprobleme, die zur Sicherstellung einer intuitiven Benutzung beseitigt werden müssen. Diese realen Probleme wurden aus 268 von 304 kritischen Ereignissen (88.16 %) abgeleitet. Die restlichen 21 der 304 kritischen Ereignisse (6.91 %) stellten lediglich (zeitliche) Ineffizienzen im Sinne der handlungsorientierten Fehlertaxonomie dar und bildeten dadurch ebenfalls keine realen Nutzungsprobleme ab. Bei der Einteilung der Nutzungsprobleme in die Fehlerkategorien der handlungsorientierten Fehlertaxonomie von Zapf et al. (1989) konnte ein Cohens κ von .78 festgestellt werden, was laut Landis und Koch (1977) als eine beachtliche Übereinstimmung interpretiert werden kann.

Zusammenfassend kann festgehalten werden, dass die Trefferrate von IntuiBeat-F bei wenig strikter Anwendung bezogen auf alle Nutzungsprobleme damit bei 93.09 % und die Fehlalarmrate bei 6.91 % lag. Betrachtet man nur die aus kritischen Ereignissen abgeleiteten realen Nutzungsprobleme, reduziert sich die Trefferrate bei wenig strikter Anwendung entsprechend auf 88.16 % und die Fehlalarmrate steigt auf 11.84 % (siehe Tabelle 6.2).

Explorative Datenanalyse von IntuiBeat-F (strikte Anwendung)

Betrachtet man ausschließlich die mithilfe von Rhythmus-Peaks erkannten kritischen Ereignisse (d.h. strikte Anwendung von IntuiBeat-F), wurden insgesamt 160 Rhythmus-Peaks generiert, wobei sich entsprechend 94 Rhythmus-Peaks (58.75 %) auf 41 Nutzungsprobleme bzw. 90 Rhythmus-Peaks (56.25 %) auf 40 reale Nutzungsprobleme verteilten. Vier der 160 Rhythmus-Peaks (2.5 %) wiesen damit auf einen Bewegungsfehler hin. Die restlichen 66 von 160 Rhythmus-Peaks (41.25 %) gingen dementsprechend nicht in Nutzungsprobleme ein und führten auch nicht zur Entdeckung von (zeitlichen) Ineffizienzen im Sinne der handlungsorientierten Fehlertaxonomie. Betrachtet man ausschließlich die durch jeweils einen Rhythmus-Peak erkannten 94 der 304 (30.92 %) insgesamt mit IntuiBeat-F gefundenen kritischen Ereignisse, gingen hierfür alle mithilfe von Rhythmus-Peaks identifizierten kritischen Ereignisse (100 %) in insgesamt 41 der insgesamt 60 Nutzungsprobleme ein (68.33 %). Beschränkt man sich nur auf die mit IntuiBeat-F abgeleiteten realen Nutzungsprobleme, konnten 90 der 94 insgesamt mithilfe von Rhythmus-Peaks identifizierten, auf reale Nutzungsprobleme hinweisenden, kritischen Ereignisse (95.75 %) 40 der 41 mithilfe von Rhythmus-Peaks abgeleiteten Nutzungsprobleme (97.56 %) identifizieren. Schließlich gingen vier der 94 (4.26 %) insgesamt mithilfe von Rhythmus-Peaks identifizierten kritischen Ereignisse in ein Nutzungsproblem ein, welches einen Bewegungsfehler als Ursache hatte. Dementsprechend lag die Trefferrate bei strikter Anwendung von IntuiBeat-F bei der Ableitung aller Nutzungsprobleme bei 100 % und die Fehlalarmrate bei 0 %, was einer Verbesserung von rund 7 % gegenüber der wenig strikten Anwendung entsprach. Betrachtet man nur die durch die strikte Anwendung von IntuiBeat-F abgeleiteten realen Nutzungsprobleme kann eine Trefferrate von 95.75 % und einer Fehlalarmrate von 4.25 % festgestellt werden, was sogar einer Verbesserung von knapp 8 % gegenüber der wenig strikten Anwendung entsprach (siehe Tabelle 6.2).

Zusammenfassend kann also festgehalten werden, dass vom Evaluator lediglich 58.75 % aller 160 Rhythmus-Peaks für die Ableitung von Nutzungsproblemen berücksichtigt und damit 41.25 % der Rhythmus-Peaks ignoriert wurden. Betrachtet man lediglich reale Nutzungsprobleme, wurden entsprechend nur 56.25 % aller 160 Rhythmus-Peaks für die Ableitung von realen Nutzungsproblemen genutzt und damit sogar 43.75 % der 160 Rhythmus-Peaks vernachlässigt. In diesem Zusammenhang wurden lediglich 30.92 % von den 304 kritischen Ereignissen mithilfe der 94 Rhythmus-Peaks entdeckt, was bezogen auf die 283 auf Nutzungsprobleme hinweisenden kritischen Ereignisse 33.22 % (94 Rhythmus-Peaks) und bezogen auf die 268, auf reale Nutzungsprobleme hinweisenden, kritischen Ereignisse 33.58 % (90 Rhythmus-Peaks) darstellte. Dementsprechend wurden 69.08 % aller kritischen Ereignisse unabhängig von Rhythmus-Peaks protokolliert. Bezogen auf alle Nutzungsprobleme wurden damit 66.78 % und bezogen auf alle realen Nutzungsprobleme 66.42 % der kritischen Ereignisse unabhängig festgehalten. Aus diesem Grund konnten bei der ausschließlichen Berücksichtigung von Rhythmus-Peaks (strikte Anwendung von IntuiBeat-F) bei IntuiBeat-F 19 Nutzungsprobleme (d.h. nur 68.33 % noch auffindbar) und 14 reale Nutzungsprobleme weniger (d.h. nur 71.43 % noch auffindbar) bei gleichzeitiger Steigerung der Trefferrate bzw. Senkung der Fehlalarmrate gegenüber der wenig strikten Anwendung von IntuiBeat-F beobachtet werden (siehe Tabelle 6.2).

Explorative Datenanalyse des Nutzertests mit retrospektivem Think-Aloud-Protokoll

Mithilfe des Nutzertests mit retrospektivem Think-Aloud-Protokoll konnten insgesamt 247 kritische Ereignisse erkannt werden, wovon unter Berücksichtigung der in der Abbildung 6.2 in Teilabschnitt 6.1.3 beschriebenen Arbeitsschritte zur Ableitung von Nutzungsproblemen (Burmester, 2016) und der handlungsorientierten Fehlertaxonomie (Zapf et al., 1989) 218 kritische Ereignisse (88.26 %) zu insgesamt 54 einzigartigen Nutzungsproblemen konsolidiert wurden, wovon drei kritische Ereignisse (1.38 %) einen Bewegungsfehler als Ursache hatten (5.56 %) und es sich damit nur bei 51 der 54 Nutzungsprobleme (94.44 %) um reale Nutzungsprobleme handelte. Die restlichen 29 kritischen Ereignisse (11.74 %) stellten lediglich (zeitliche) Ineffizienzen im Sinne der handlungsorientierten Fehlertaxonomie und damit auch keine echten Nutzungsprobleme im Bezug auf intuitive Benutzung dar. Dementsprechend ließen sich 51 reale Nutzungsprobleme aus 215 der 247 kritischen Ereignisse ableiten (87.05 %).

Zusammenfassend kann festgehalten werden, dass die Trefferrate bezogen auf alle aus kritischen Ereignissen abgeleiteten Nutzungsprobleme geringer als bei IntuiBeat-F bei 88.26 % und die Fehlalarmrate höher als bei IntuiBeat-F bei 11.74 % lag. Betrachtet man nur die aus kritischen Ereignissen abgeleiteten realen Nutzungsprobleme reduziert sich die Trefferrate entsprechend auf 87.05 % und die Fehlalarmrate steigt auf 12.95 % (siehe Tabelle 6.2).

Qualitative Ursachenanalyse mit beiden Methoden gefundener Nutzungsprobleme bei wenig strikter Anwendung von IntuiBeat-F

Beurteilt man die, mithilfe von IntuiBeat-F (d.h. weniger strikte Anwendung) und dem Nutzertest mit retrospektivem Think-Aloud-Protokoll abgeleiteten, Nutzungsprobleme mit der handlungsorientierten Fehlertaxonomie qualitativ, stellt man bei Berücksichtigung aller kritischen Ereignisse folgende Einteilung fest (siehe Abbildung 6.5). Im Falle von IntuiBeat-F konnten ein Urteilsfehler (12.50 %), fünf Erkennensfehler (62.50 %), ein Denkfehler (12.50 %) und ein Bewegungsfehler (12.50 %) als Ursachen für die acht Nutzungsprobleme klassifiziert werden. Damit überwogen Ursachen, die der perzeptiv-begrifflichen Ebene zugeschrieben werden können. Bezüglich des Nutzertests mit retrospektivem Think-Aloud-Protokoll konnten ein Urteilsfehler (50 %) und ein Erkennensfehler (50 %) als Ursachen für die beiden Nutzungsprobleme klassifiziert werden. Die Aufteilung der Ursachen war damit auf intellektueller und perzeptiv-begrifflicher Regulationsebene ausgeglichen. Die Ursachen für die von beiden Methoden unter Berücksichtigung aller kritischen Ereignisse abgeleiteten 52 Nutzungsprobleme waren zwölf Gewohnheitsfehler (23.08 %), zwölf Urteilsfehler (23.08 %), ein Wissensfehler (1.92 %), ein Merk-/Vergessensfehler (1.92 %), drei Unterlassensfehler (5.77 %), 16 Erkennensfehler (30.77 %), vier Denkfehler (7.69 %) und drei Bewegungsfehler (5.77 %). Die von beiden Methoden gefundenen Ursachen teilen sich dementsprechend mit 1.92 % auf die Regulationsgrundlage, mit 32.69 % auf die intellektuelle, mit 59.62 % auf die perzeptiv-begriffliche und mit 5.77 % auf die sensomotorische Regulationsebene auf.

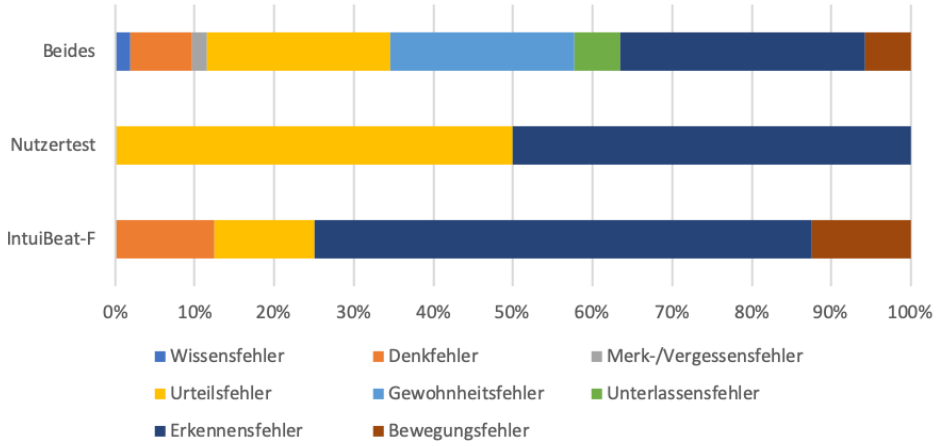


Abbildung 6.5. Ursachenklassifikation mithilfe der handlungsorientierten Fehlertaxonomie (Zapf, Brodbeck, & Prümper, 1989) der mit IntuiBeat-F (weniger strikte Anwendung) und einem Nutzertest mit retrospektivem Think-Aloud-Protokoll abgeleiteten Nutzungsprobleme im Rahmen des vierten Experiments.

Qualitative Ursachenanalyse mit beiden Methoden gefundener Nutzungsprobleme bei strikter Anwendung von IntuiBeat-F

Berücksichtigt man bei dieser qualitativen Beurteilung bei IntuiBeat-F ausschließlich die mithilfe des Algorithmus abgeleiteten Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F), ergibt sich folgende geänderte Einteilung (siehe Abbildung 6.6). Im Falle von IntuiBeat-F konnten drei Erkennensfehler (75 %) und ein Bewegungsfehler (25 %) als Ursachen für die vier Nutzungsprobleme klassifiziert werden, wodurch auch hier die Ursachen hauptsächlich der perzeptiv-begrifflichen Ebene zugeordnet werden können. Bezüglich des Nutzertests mit retrospektivem Think-Aloud-Protokoll konnten ein Gewohnheitsfehler (5.88 %), vier Urteilsfehler (23.53 %), ein Wissensfehler (5.88 %), ein Merk-/Vergessensfehler (5.88 %), zwei Unterlassensfehler (11.77 %), vier Erkennensfehler (23.53 %), ein Denkfehler (5.88 %) und drei Bewegungsfehler (17.65 %) als Ursachen für die 17 Nutzungsfehler identifiziert werden, wodurch sich auch hier die zugrunde liegenden Ursachen hauptsächlich der perzeptiv-begrifflichen Ebene zuordnen lassen. Die Ursachen für die, von beiden Methoden unter Berücksichtigung nur der bei IntuiBeat-F algorithmisch abgeleiteten, 37 Nutzungsprobleme waren elf Gewohnheitsfehler (29.73 %), neun Urteilsfehler (24.32 %), ein Unterlassensfehler (2.70 %), 13 Erkennensfehler (35.14 %) und drei Denkfehler (8.11 %). Die von beiden Methoden gefundenen Ursachen teilen sich dementsprechend mit 32.43 % auf die intellektuelle und mit mit 67.57 % auf die perzeptiv-begriffliche Regulationsebene auf.

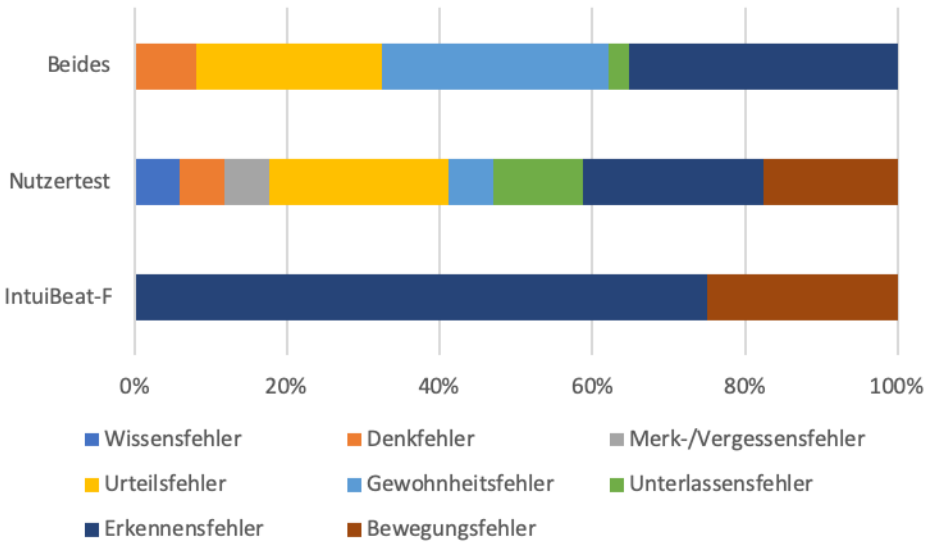


Abbildung 6.6. Ursachenklassifikation mithilfe der handlungsorientierten Fehlertaxonomie (Zapf, Brodbeck, & Prümper, 1989) der mit IntuiBeat-F (strikte Anwendung) und einem Nutzertest mit retrospektivem Think-Aloud-Protokoll abgeleiteten Nutzungsprobleme im Rahmen des vierten Experiments.

6.1.5 Diskussion

Im vorliegenden vierten Experiment wurde die wissenschaftliche Güte von IntuiBeat-F bezüglich der Gütekriterien *Gründlichkeit* und *Gültigkeit*, sowie der *zeitlichen Anwendungseffizienz*, als für das Projekt 3D-GUIde wichtiges Kriterium der praktischen Güte, beim Vergleich mit einem als Quasi-Außenkriterium fungierenden Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der der Nutzung der CUI-Regalplanungs-Software *IPO.Rack* geprüft. Bei dieser Software handelte es sich auf Basis von Experteneinschätzungen und Nutzereinschätzungen (d.h. anhand eines TFQ) um eine eher wenig intuitiv benutzbare Software, bei der es entsprechend zu vielen, intuitive Benutzung beeinträchtigenden, Nutzungsproblemen (d.h. reale Nutzungsprobleme) kommen sollte.

Anhand der statistischen Tests wurde zunächst überprüft, ob sich ein Unterschied bezüglich der Gründlichkeit zu Gunsten von IntuiBeat-F feststellen lässt (Hypothese H1). Anschließend wurde untersucht, ob auch bezüglich der Gültigkeit ein Unterschied zu Gunsten von IntuiBeat-F erkennbar ist (Hypothese H2). Abschließend wurde überprüft, inwiefern IntuiBeat-F eine höhere zeitliche Anwendungseffizienz als ein Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der formativen Evaluation intuitiver Benutzung der Regalplanungssoftware *IPO.Rack* aufweisen kann (Hypothese H3). Im folgenden Verlauf werden die Hypothesen bezüglich der Gründlichkeit, der Gültigkeit und der zeitlichen Anwendungseffizienz unter Berücksichtigung der festgestellten Ergebnisse diskutiert, nachdem im Vorfeld auf eine mögliche Konfundierung durch die *Vorerfahrung bei der Nutzung von CUIs* eingegangen wurde.

6.1.5.1 Überprüfung einer möglichen Konfundierung durch die Vorerfahrung bei der Nutzung von CUIs

Die Vorerfahrung bei der Nutzung von CUIs unterschied sich erwartungsgemäß nicht signifikant zwischen den beiden Ausprägungen der unabhängigen Variable *Art der formativen Evaluationsmethode*. Jedoch war die Teststärke aufgrund des geringen beobachteten Effekts und der kleinen Stichprobe gering. Wie bereits zuvor beim ersten Experiment (siehe Abschnitt 5.1) angesprochen, könnte eine mögliche Erklärung für dieses Ergebnis darin bestehen, dass bei den als Stichprobe genutzten Studierenden der Medienkommunikation und der Mensch-Computer-Systeme keine großen Unterschiede bezüglich der Vorerfahrung mit CUIs zu erwarten sind (d.h. aufgrund der nicht vorhandenen Anforderungen von CUI-Kenntnissen und der allgemeinen Studieninhalte, der in der Stichprobe berücksichtigten Studiengänge, sind bezüglich der Vorerfahrung keine größeren Schwankungen zu erwarten und diese ähnlich gering ausgeprägt).

Dementsprechend ist die praktische Bedeutsamkeit der in diesem Zusammenhang ermittelten Effekte bzw. der Erkenntnisgewinn aufgrund der verwendeten Stichprobe eher als unbedeutend einzustufen (siehe Döring & Bortz, 2016), was sich auch in den sehr geringen deskriptiven Unterschieden erkennen lässt. Da darüber hinaus bei der Kontrollvariable *Vorerfahrung bei der Nutzung von CUIs* keine signifikanten Korrelationen bezüglich der Gründlichkeit, der Gültigkeit und der zeitlichen Anwendungseffizienz bei beiden formativen Evaluationsmethoden vorlagen ($p > .05$), lag mit hoher Wahrscheinlichkeit keine Konfundierung durch eine unterschiedliche Vorerfahrung bei der Nutzung von CUIs vor.

6.1.5.2 Überprüfung der Gründlichkeit

Unerwarteterweise zeigte sich bei der Evaluation einer weniger intuitiv benutzbaren Software die prognostizierte höhere Gründlichkeit von IntuiBeat-F gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll inferenzstatistisch lediglich bei Berücksichtigung aller Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F), und nicht bei ausschließlicher Berücksichtigung der durch Rhythmus-Peaks entdeckter Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F). Bei Berücksichtigung aller Nutzungsprobleme konnte IntuiBeat-F gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll eine signifikant höhere Gründlichkeit von 32 % im Vergleich zu 22 % zeigen (siehe Teilabschnitt 6.1.4.3). Betrachtet man diese Ergebnisse unter Berücksichtigung bisheriger Meta-Evaluationen aus dem Bereich der Usability-Forschung (siehe Koutsabasis et al., 2007) lässt sich feststellen, dass empirische Nutzertests (d.h. Nutzertest mit parallelem Think-Aloud-Protokoll, Nutzertest mit gemeinsamen Explorieren) zur formativen Evaluation von Gebrauchstauglichkeit (d.h. nicht speziell intuitive Benutzung) eine Gründlichkeit zwischen circa 20 % und 40 % aufweisen (d.h. variiert stark aufgrund verschiedener Faktoren wie dem Evaluatoreffekt, siehe Alshamari & Mayhew, 2009). IntuiBeat-F bei wenig strikter Anwendung liegt damit mit 32 % in der Mitte und der Nutzertest mit retrospektivem Think-Aloud-Protokoll am unteren Ende bei 22 %. Zusammenfassend lässt sich somit die in diesem Zusammenhang aufgestellte Hypothese H1.A bei der Evaluation einer weniger intuitiv benutzbaren Software bestätigen (siehe Tabelle 6.3).

Nichtsdestotrotz konnte bei der Berücksichtigung von ausschließlich mithilfe von Rhythmus-Peaks abgeleiteten realen Nutzungsproblemen (d.h. strikte Anwendung von IntuiBeat-F) kein signifikanter Unterschied zu Gunsten von IntuiBeat-F, und deskriptiv sogar eine Verschlechterung der Gründlichkeit um 13 % gegenüber der wenig strikten Anwendung festgestellt werden (siehe Teilabschnitt 6.1.4.3). Betrachtet man die deskriptiven Daten des vierten Experiments unter diesem Aspekt genauer (siehe Teilabschnitt 6.1.4.6), ist dies nicht verwunderlich, da lediglich rund 60 % aller Rhythmus-Peaks für die Identifikation von kritischen Ereignissen, die auf reale Nutzungsprobleme hinweisen konnten, genutzt, nur rund 34 % dieser Ereignisse mithilfe von Rhythmus-Peaks entdeckt und demzufolge auch rund 30 % weniger reale Nutzungsprobleme mithilfe von IntuiBeat-F abgeleitet werden konnten. Eine deskriptive Verschlechterung der Gründlichkeit von IntuiBeat-F um 13 % bei gleichzeitiger Verbesserung der Gründlichkeit beim Einsatz des Nutzertests mit retrospektivem Think-Aloud-Protokoll um 1 % ist die Folge (siehe Teilabschnitt 6.1.4.3), da der Nutzertest nun im Vergleich zu IntuiBeat-F einen größeren Anteil aller real verfügbaren Nutzungsprobleme im System identifiziert (d.h. es fallen hier keine realen Nutzungsprobleme weg). Da die Gesamtzahl aller vorhandenen realen Nutzungsprobleme durch eine Vereinigungsmenge der durch beide Evaluationsmethoden gefundenen Nutzungsprobleme bestimmt wurde, profitiert die Gründlichkeit des Nutzertests mit retrospektivem Think-Aloud-Protokoll von dieser Tatsache (siehe Teilabschnitt 6.1.3.3).

Eine mögliche Erklärung für die deskriptive Verschlechterung der Gründlichkeit bei der strikten gegenüber der weniger strikten Anwendung von IntuiBeat-F liegt dabei in der Tatsache, dass der Evaluator, statt die übrigen 40 % der Rhythmus-Peaks für die Identifikation von kritischen Ereignissen zu nutzen, einen großen Anteil an kritischen Ereignissen unabhängig von Rhythmus-Peaks protokolliert hat, aus denen sich letztendlich ebenfalls

reale Nutzungsprobleme ableiten konnten. Dementsprechend ergibt sich die Frage, ob die knapp 70 % der unabhängig vom Algorithmus protokollierten kritischen Ereignisse, die auf reale Nutzungsprobleme hinweisen konnten, auch bei strikter Anwendung von IntuiBeat-F entdeckbar gewesen wären, wenn der Evaluator die restlichen knapp 40% der Rhythmus-Peaks berücksichtigt hätte (siehe Teilabschnitt 6.1.4.6).

Falls der Evaluator die Rhythmus-Peaks einfach vorsätzlich ignoriert hat, weil ihm die Zuordnung der Rhythmus-Peaks zur entsprechenden Videoposition mithilfe der Marker-CSV-Datei während der retrospektiven Gesprächsführung schwer fiel, sollte in künftigen Experimenten der Ablauf des retrospektiven Interviews überdacht, und dem Evaluator eine einfachere Zuordnungsmöglichkeit bereitgestellt werden. Diese sollte am besten direkt in die Videoaufzeichnung integrierbar sein. Die eben vorgeschlagene Erklärung kann als durchaus wahrscheinlich eingestuft werden, da es sich bei dem Evaluator des vierten Experiments um einen Bachelorstudenten handelte, der zwar eine hohe Kenntnis des Untersuchungsgegenstands hatte, aber zuvor lediglich eine weitere Nutzerteststudie durchführte. Das zeitlich aufwendige Zuordnen der Rhythmus-Peaks zu den entsprechenden Videostellen könnte ihn dementsprechend aufgrund der beschränkten Zeit im retrospektivem Interview leicht überfordert haben, weswegen er sich aufgrund der hohen Umständlichkeit stattdessen kritische Ereignisse unabhängig vom Algorithmus notiert haben könnte. Diese mögliche Erklärung erscheint nachvollziehbar und liefert einen möglichen Grund für die versäumten kritischen Ereignisse, sowie die damit verbundene Verschlechterung der Gründlichkeit bei strikter Anwendung von IntuiBeat-F.

Sollte diese unberücksichtigten Rhythmus-Peaks echte Fehlalarme (d.h. hinter diesen verbirgt sich kein kritisches Ereignis) und nicht die Fehlinterpretation des Evaluators darstellen, kann diese Fehlalarmrate im Vergleich mit Studien aus dem CUI-Bereich als eher hoch eingestuft werden. Beispielsweise konnten Akers et al. (2012) bei der Meta-Evaluation ihrer formativen Evaluationsmethode für Gebrauchstauglichkeit (d.h. nicht speziell intuitive Benutzung), die zur Identifikation von Nutzungsproblemen algorithmisch ermittelte Backtracking-Operationen nutzt (d.h. Undo- und Löschoperationen als Indikatoren für kritische Ereignisse), eine Fehlalarmrate von circa 30 % bei der Evaluation von Google SketchUp und eine Fehlalarmrate von circa 5 % bei der Evaluation von Adobe Photoshop CS3 feststellen. Jedoch zeigten die Ergebnisse von Akers et al. (2012) gleichzeitig, dass bereits bei zwei Studien eine riesige Bandbreite in der Fehlalarmrate bei algorithmisch ermittelten Ereignisindikatoren und eine hohe Variabilität in Abhängigkeit vom untersuchten System bestehen kann, weswegen eine konzeptuelle Replikation des vierten Experiments mit einem anderen System als Untersuchungsgegenstand ratsam wäre. Dadurch könnte besser verstanden werden, warum der prognostizierte Unterschied zwischen den beiden formativen Evaluationsmethoden nicht beobachtbar war.

Als Untersuchungsgegenstand würde sich eine stärker intuitiv benutzbare Software eignen, da diese mit hoher Wahrscheinlichkeit weniger (reale) Nutzungsprobleme aufweist, derselbe Evaluator damit weniger bei der Zuordnung von Rhythmus-Peaks überfordert wäre, und dabei auch gleichzeitig die Gründlichkeit des Evaluators grundsätzlich beim Einsatz beider formativer Evaluationsmethoden höher sein sollte. An dieser Stelle ist außerdem anzumerken, dass der Evaluator sich im Rahmen des vierten Experiments lediglich notiert hat, welcher Rhythmus-Peak zu welchen kritischen Ereignissen führte, aber nicht anmerkte, ob er diesen vorsätzlich ignoriert hat oder dieser nutzlos war. Wegen der bereits anstrengenden

manuellen Zuordnung von Rhythmus-Peaks zu Videostellen wurde auf diese zusätzliche Fehlerquelle verzichtet, da die Qualität der Nutzungsprobleme für das Anwenderprojekt entscheidend war. Eine solche Aktion würde außerdem mit hoher Wahrscheinlichkeit zu Lasten der zeitlichen Anwendungseffizienz gehen und einen Vergleich mit dem Nutzertest mit retrospektivem Think-Aloud-Protokoll verzerren.

Zusammenfassend kann IntuiBeat-F bei strikter Anwendung somit unabhängig vom Grund für die versäumten kritischen Ereignisse keine Gründlichkeit bei der Evaluation einer weniger intuitiv benutzbaren Software attestiert werden, wenn bei IntuiBeat-F ausschließlich Nutzungsprobleme bei der Bestimmung der Gründlichkeit, die auf Rhythmus-Peaks rückführbar sind, berücksichtigt wurden. Aus diesem Grund muss auch die in diesem Zusammenhang aufgestellte Hypothese H1.B verworfen werden (siehe Tabelle 6.3). Aufgrund der Tatsache, dass zwar mit hoher Wahrscheinlichkeit die aufwendige manuelle Zuordnung für die Verschlechterung der Gründlichkeit bei strikter Anwendung von IntuiBeat-F verantwortlich gemacht werden kann, man aber zunächst verifizieren möchte, ob ähnliche Befunde auch in einer Umgebung auftreten, in der ein Novize als Evaluator genug Zeit für die manuelle Zuordnung aller Rhythmus-Peaks zur Verfügung haben sollte (d.h. mehr Zeit, da wahrscheinlich weniger Nutzungsprobleme auftreten werden), soll in einem Folgeexperiment zunächst eine konzeptuelle Replikation mit einer stärker intuitiv benutzbaren Software erfolgen.

6.1.5.3 Überprüfung der Gültigkeit

Wie erwartet, zeigte sich bei der Evaluation einer weniger intuitiv benutzbaren Software signifikant eine höhere Gültigkeit von IntuiBeat-F gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll, sowohl bei Berücksichtigung aller Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) als auch bei Berücksichtigung der ausschließlich mithilfe von Rhythmus-Peaks entdeckten Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F). In beiden Fällen konnte bei IntuiBeat-F eine Gültigkeit von über 90 % (d.h. weniger strikt: 92 %; strikt: 96 %) und beim Nutzertest mit retrospektivem Think-Aloud-Protokoll eine Gültigkeit von 81 % beobachtet werden, was im Vergleich zu anderen formativen Evaluationsmethoden innerhalb der Usability-Forschung (d.h. Evaluationsmethoden die Gebrauchstauglichkeit formativ messen, aber nicht speziell intuitive Benutzung) in einem ähnlich hohen Wertebereich (d.h. zwischen 74 und 90 % bei empirischen Nutzertests) liegt (siehe Koutsabasis et al., 2007). An dieser Stelle ist bezüglich der Studie von Koutsabasis et al. (2007) anzumerken, dass die Evaluatorenteams für die dort verglichenen vier formativen Evaluationsmethoden (d.h. zwei analytische Methoden: heuristische Evaluation, Cognitive Walkthrough; zwei empirische Methoden: Nutzertest mit parallelem Think-Aloud-Protokoll, Nutzertest mit gemeinsamen Explorieren) überwiegend geringe Vorerfahrung mit der formativen Evaluation besaßen, es sich jedoch dabei wie in dieser Studie auch um Studierende mit hoher Kenntnis des Untersuchungsgegenstandes und dementsprechend um Novizen auf dem Gebiet handelte.

Nichtsdestotrotz lässt sich bei der Berücksichtigung ausschließlich durch Rhythmus-Peaks identifizierter Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F) gegenüber der Berücksichtigung aller Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) deskriptiv eine Verbesserung der Gültigkeit von rund 4 % bei IntuiBeat-F feststellen, obwohl mithilfe der Rhythmus-Peaks nur circa 68 % aller Nutzungsprobleme und 72 % aller realen Nutzungsprobleme identifiziert werden konnten (siehe Teilabschnitt 6.1.4.6). Aufgrund der Tatsache, dass jedoch mithilfe der Rhythmus-Peaks nur 41 aller 60 Nutzungsprobleme identifiziert werden konnten (also 31.67 % weniger als bei der weniger strikten Anwendung) bzw. 40 aller 56 realen Nutzungsprobleme (also 28.57 % weniger als bei der weniger strikten Anwendung), hat sich die Gültigkeit bei strikter Anwendung verbessert, da diese ja das Verhältnis von realen zu allen gefundenen Nutzungsproblemen abbildet. Es wurden dementsprechend mehr nicht reale Nutzungsprobleme nicht berücksichtigt als reale Nutzungsprobleme (d.h. Verhältnis von 56 zu 60 insgesamt; verbessertes Verhältnis von 40 zu 41 durch Rhythmus-Peaks). Wie bereits im Rahmen der Diskussion der Gründlichkeit im vorangegangen Abschnitt erwähnt, sollte jedoch anhand einer konzeptuellen Replikation mit einer stärker intuitiv benutzbaren Software zunächst verifiziert werden, inwiefern der Evaluator Rhythmus-Peaks vorsätzlich aufgrund der komplizierten manuellen Zuordnung von Rhythmus-Peaks zu den entsprechenden Videostellen vergessen hat und inwieweit sich neben der Gründlichkeit auch die Gültigkeit von IntuiBeat-F durch Berücksichtigung der übrigen 60 % aller verfügbaren Rhythmus-Peaks systematisch verbessern kann.

Zusammenfassend können jedoch unabhängig von diesen Schlussfolgerungen, die Ergebnisse des vorliegenden vierten Experiments bezüglich der Gültigkeit bei der Evaluation einer weniger intuitiv benutzbaren Software als beeindruckend interpretiert und die in diesem Zusammenhang aufgestellten Hypothesen H2.A und H2.B entsprechend als bestätigt angesehen werden (siehe Tabelle 6.3). Da zwar bereits bei manueller Zuordnung der Rhythmus-Peaks zu den kritischen Ereignissen gezeigt werden konnte, dass durch die strikte Anwendung von IntuiBeat-F die Gültigkeit der Ergebnisse bei einer weniger intuitiv benutzbaren Software im Vergleich zur wenig strikten Anwendung verbessert werden kann, soll in einem Folgeexperiment zunächst verifiziert werden, ob dies auch für eine stärker intuitiv benutzbare Software gilt, in der ein Novize als Evaluator mit hoher Wahrscheinlichkeit mehr Zeit für die manuelle Zuordnung zur Verfügung hat und dementsprechend grundsätzlich weniger reale Nutzungsprobleme vergessen kann (d.h. mehr Zeit, da wahrscheinlich weniger Nutzungsprobleme auftreten werden).

6.1.5.4 Überprüfung der zeitlichen Anwendungseffizienz

Wie erwartet, zeigte sich signifikant eine höhere zeitliche Anwendungseffizienz von IntuiBeat-F gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der Evaluation einer weniger intuitiv benutzbaren Software. IntuiBeat-F nahm hier sogar 20 % weniger Zeit für die Anwendung in Anspruch (siehe Teilabschnitt 6.1.4.5). In einer vergleichbaren Studie aus dem CUI-Bereich, in der Akers et al. (2012) eine neu entwickelte formative Evaluationsmethode für Gebrauchstauglichkeit (d.h. nicht speziell intuitive Benutzung), welche die Identifikation von Nutzungsproblemen auf Basis kritischer Ereignisse mithilfe

von algorithmisch ermittelten Backtracking-Operationen (d.h. Undo- und Löschoperationen als Indikatoren für kritische Ereignisse) vornimmt, mit einem gewöhnlichen Nutzertest hinsichtlich der zeitlichen Anwendungseffizienz verglichen, konnte ebenfalls eine Verbesserung um circa 20 % durch die erhöhte Automatisierung festgestellt werden. Akers et al. (2012) verwendeten jedoch anstelle eines retrospektiven Think-Aloud-Protokolls ein paralleles Think-Aloud-Protokoll als Benchmark. Dementsprechend sind die Ergebnisse dieses vierten Experiments noch beeindruckender, da das von ihm verwendete parallele Think-Aloud-Protokoll nicht zusätzlich, die für das retrospektive Interview erforderliche Moderation des Evaluators in die Bestimmung der zeitlichen Anwendungseffizienz miteinbezieht.

Darüber hinaus kann mit hoher Wahrscheinlichkeit aufgrund der Tatsache, dass vom Evaluator lediglich rund 60 % aller Rhythmus-Peaks bei der Durchführung des retrospektiven Interviews für die Identifikation von realen Nutzungsproblemen verwendet wurden und der Evaluator stattdessen circa 66 % aller kritischen Ereignisse, die auf reale Nutzungsprobleme hinweisen konnten, unabhängig von Rhythmus-Peaks entdeckte (siehe Teilabschnitt 6.1.4.6), mit einer noch zeitlich effizienteren Durchführung gerechnet werden, wenn ein höherer Anteil an Rhythmus-Peaks beachtet wird und diese auch zu realen Nutzungsproblemen führen. Dementsprechend kann die in diesem Zusammenhang aufgestellte Hypothese H3 bei der Evaluation einer weniger intuitiv benutzbaren Software als bestätigt angesehen werden (siehe Tabelle 6.3). Wie bereits im Rahmen der Diskussionen zur Gründlichkeit und Gültigkeit angesprochen, soll dieses Ergebnis zunächst in einem Folgeexperiment anhand einer konzeptuellen Replikation mit einer stärker intuitiv benutzbaren Software verifiziert werden.

6.1.6 Schlussfolgerung

Tabelle 6.3. *Übersicht der mithilfe des vierten Experiments bestätigten Hypothesen im Zuge der Meta-Evaluation von IntuiBeat-F (KEs: kritische Ereignisse; RP-KEs: mit Rhythmus-Peaks assoziierte kritische Ereignisse).*

Hypothese	Experiment 4
(H1) Überprüfung der Gründlichkeit:	
- (A) Alle KEs	✓
- (B) Nur RP-KEs	✗
(H2) Überprüfung der Gültigkeit:	
- (A) Alle KEs	✓
- (B) Nur RP-KEs	✓
(H3) Überprüfung der zeitlichen Anwendungseffizienz	✓

Zusammenfassend kann festgehalten werden, dass die wissenschaftliche Güte von IntuiBeat-F als formative Evaluationsmethode für intuitive Benutzung hinsichtlich des Gütekriteriums *Gültigkeit* bei der Evaluation einer weniger intuitiv benutzbaren Software empirisch sowohl bei strikter als auch wenig strikter Anwendung bestätigt werden konnte (siehe

Tabelle 6.3). Außerdem konnte IntuiBeat-F auch praktische Güte aufgrund seiner, im Vergleich zum Nutzertest mit retrospektivem Think-Aloud-Protokoll, höheren zeitlichen Anwendungseffizienz empirisch bestätigt werden (siehe Tabelle 6.3). Konfundierungen durch Unterschiede in der Vorerfahrung bei der Nutzung von CUIs können mit hoher Wahrscheinlichkeit ausgeschlossen werden. Die Effekt- und Teststärken lagen dabei überwiegend im oberen Bereich. Bezüglich des wissenschaftlichen Gütekriteriums der Gründlichkeit konnte die wissenschaftliche Güte von IntuiBeat-F nicht bestätigt werden, wenn IntuiBeat-F vom Evaluator strikt angewendet und deswegen nur Nutzungsprobleme als Grundlage für die Bestimmung der Gründlichkeit genutzt wurden, die mithilfe von Rhythmus-Peaks gefunden wurden. Dementsprechend führte der Einsatz von Rhythmus-Peaks (d.h. strikte Anwendung von IntuiBeat-F) bei der Evaluation einer weniger intuitiv benutzbaren Software nicht automatisch zu einer signifikant höheren Gründlichkeit gegenüber einem Nutzertest mit retrospektivem Think-Aloud-Protokoll, sondern nur wenn der Evaluator unabhängig vom Algorithmus selbst noch reale Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) identifizieren konnte (siehe Tabelle 6.3).

Wie eben in der Diskussion angesprochen, gelang die manuelle Zuordnung der Rhythmus-Peaks zu den Videostellen während des retrospektiven Interviews dem Evaluator des vierten Experiments nicht vollständig, weswegen, für ein besseres Verständnis der Hintergründe der versäumten kritischen Ereignisse, als nächstes eine konzeptuelle Replikation des vierten Experiments mit einer stärker intuitiv gestalteten Software als formativer Untersuchungsgegenstand durchgeführt werden soll. Die Verwendung einer stärker intuitiv benutzbaren Software, im Vergleich zu der in diesem Experiment verwendeten weniger intuitiv benutzbaren Software, sollte allgemein zu weniger Nutzungsproblemen führen, und dem Evaluator trotz Gesprächsführung ermöglichen, besser auf die Rhythmus-Peaks zu achten. Bei gleichbleibender Vorgehensweise und Beibehaltung desselben Evaluators kann so außerdem verifiziert werden, ob der Evaluator auch unter diesen Umständen einen ähnlichen Anteil an Rhythmus-Peaks vorsätzlich ignoriert oder ob die Fehlalarmrate von IntuiBeat-F wirklich bei 40 % liegt. Mithilfe der folgenden konzeptuellen Replikation sollen offene Fragen in diesem Zusammenhang geklärt und Optimierungen für weitere Folgeexperimente abgeleitet werden, um die wissenschaftliche Güte und zeitliche Anwendungseffizienz von IntuiBeat-F zukünftig durch eine stärkere Unterstützung des Evaluators bei der Zuordnung von Rhythmus-Peaks zu Videostellen verbessern zu können.

6.2 Experiment 5

Da das vorige Experiment daran scheiterte, die wissenschaftliche Güte von IntuiBeat-F hinsichtlich des formalen Hauptgütekriteriums *Gründlichkeit* vollständig nachzuweisen (d.h. Hypothese H1.B verletzt und damit das Gütekriterium der Gründlichkeit nicht bei strikter Anwendung von IntuiBeat-F nachweisbar, siehe Tabelle 6.3), untersuchte das, in diesem Abschnitt vorgestellte, Experiment aus den in der Diskussion des vierten Experiments genannten Gründen (siehe Teilabschnitt 6.1.5), ob die Ergebnisse des vierten Experiments konzeptuell replizierbar sind, wenn als Untersuchungsgegenstand für die Meta-Evaluation die Hotelbuchungswebseite *HolidayCheck.de* (HolidayCheck AG, 2017) zum Einsatz kommt, die auf Basis einer Experteneinschätzung als eine stärker intuitiv benutzbare Software eingestuft werden kann. Es kam dementsprechend wie im Vorgänger-

experiment zu einem Vergleich von IntuiBeat-F mit einem als Quasi-Außenkriterium fungierenden Nutzertest mit retrospektivem Think-Aloud-Protokoll. Das fünfte Experiment betrachtete dementsprechend auch die zweite Forschungsfrage und den formativen Aspekt der dritten Forschungsfrage dieser Arbeit.

6.2.1 Überprüfung der Gründlichkeit und Gültigkeit von IntuiBeat-F

Aufgrund der Tatsache, dass es sich beim fünften Experiment um eine konzeptuelle Replikation des vierten Experiments unter Berücksichtigung einer neuen Domäne bzw. einer stärker intuitiv benutzbaren Software handelte, wurden hier auch die gleichen Hypothesen wie im Vorgängerexperiment aufgestellt, weswegen an dieser Stelle auf den entsprechenden Teilabschnitt 6.1.1 des vierten Experiments verwiesen werden soll.

6.2.2 Überprüfung der zeitlichen Anwendungseffizienz von IntuiBeat-F

Auch bezüglich der zeitlichen Anwendungseffizienz wurde die entsprechende Hypothese aus dem Vorgängerexperiment übernommen, weswegen an dieser Stelle auf den entsprechenden Teilabschnitt 6.1.2 des vierten Experiments verwiesen werden soll.

6.2.3 Methode

6.2.3.1 Teilnehmer

Für das fünfte Experiment wurden 24 Versuchspersonen über das Probandensystem des Instituts für Mensch-Computer-Medien an der Universität Würzburg rekrutiert. Da keine Versuchsperson aufgrund unvollständiger oder fehlerhafter Daten von der Datenauswertung ausgeschlossen werden musste, konnten für die Meta-Evaluation von IntuiBeat-F 24 Versuchspersonen berücksichtigt werden, welche alle rechtsfüßig (d.h. der rechte Fuß stellte den dominanten Fuß dar) waren. Die Versuchspersonen setzten sich dabei aus 17 Frauen und sieben Männern zusammen. Das Durchschnittsalter betrug 20.67 Jahre ($SD = 2.20$). Es handelte sich bei allen Teilnehmern um Studierende der Julius-Maximilians-Universität Würzburg, wobei sieben Personen Mensch-Computer-Systeme (29.20 %) und 17 Personen Medienkommunikation (70.80 %) studierten. Alle Versuchsteilnehmer wurden über das Probanden-System des Instituts Mensch-Computer-Medien über eine gesonderte Mail darauf hingewiesen, für den Versuch flache Sportschuhe zu tragen, um eine möglichst problemlose Rhythmuseingabe über das USB-Fußpedal zu ermöglichen. Für die Teilnahme an der Untersuchung bekam jede Versuchsperson eine Versuchspersonenstunde gutgeschrieben. Die mit einem TFQ gemessene Vorerfahrung der Versuchspersonen bezüglich der Nutzung von Hotelbuchungswebseiten betrug im Durchschnitt 1.21 ($SD = .78$) bei einem Maximum von 6 und lag damit erwartungsgemäß im höheren Bereich. Alle Versuchspersonen besaßen damit eine mittelmäßige Vorerfahrung mit Hotelbuchungswebseiten. Alle Versuchspersonen gaben an, am Experiment freiwillig teilzunehmen.

6.2.3.2 Versuchsdesign

Für die Beantwortung der zweiten Forschungsfrage und des formativen Aspekts der dritten Forschungsfrage wurde das gleiche Experimentaldesign wie zuvor beim vierten Experiment genutzt. Für genauere Informationen bezüglich der unabhängigen und abhängigen Variablen wird daher auf den entsprechenden Absatz 6.1.3.1 des vierten Experiments verwiesen.

6.2.3.3 Versuchsmaterialien und Maße

Die im fünften Experiment verwendeten Versuchsmaterialien und Maße unterschieden sich nur in einigen Punkten von den im vierten Experiment verwendeten Versuchsmaterialien und Maßen, weswegen in diesem Teilabschnitt auch nur auf diese eingegangen wird, und für die übernommenen Versuchsmaterialien und Maße auf den entsprechenden Absatz 6.1.3.3 des vierten Experiments verwiesen wird.

Untersuchungsgegenstand der formativen Evaluation: HolidayCheck.de

Als Untersuchungsgegenstand für die Meta-Evaluation von IntuiBeat-F kam die Hotelbuchungswebseite *HolidayCheck.de* der HolidayCheck AG (2017) als stärker intuitiv benutzbare Software auf Basis einer qualitativen Experteneinschätzung ($N_{Experte}$= 5; Vorgehen: siehe entsprechenden Absatz innerhalb des Teilabschnitts 5.1.4.3) im Rahmen des fünften Experiments zum Einsatz (siehe Abbildung 6.7). Es wurde für das fünfte Experiment explizit kein CUI als Untersuchungsgegenstand verwendet, da die Robustheit der Ergebnisse des vierten Experiments anhand einer konzeptuellen Replikation mit einem Untersuchungsgegenstand aus einer anderen Domäne überprüft werden sollte.

Es wurde sich als Untersuchungsgegenstand für eine Hotelplanungswebseite entschieden, da zum einen unter Berücksichtigung der, für das Experiment zur Verfügung stehenden, Studierendenstichprobe zu erwarten ist, dass HolidayCheck.de von diesen überwiegend als stärker intuitiv benutzbare Softwareanwendung eingestuft wird (d.h. stärker intuitiv benutzbare Softwareanwendungen sollten logischerweise auch zu weniger Nutzungsproblemen führen), da so gut wie jeder eine derartige Webseite schon einmal genutzt hat. Zum anderen teilen sich Webseiten mit CUIs die grundlegende Gemeinsamkeit, dass der Nutzer bei der Systeminteraktion aufgrund ihrer Dynamik (z.B. diverse Nutzerziele und Lösungswege möglich) substantiell unterschiedliche Erfahrungen macht und damit auch auf verschiedene Nutzungsprobleme stößt (siehe Akers, 2010). Webseiten besitzen ebenfalls eine derartige Dynamik, da sich Nutzer üblicherweise für die Lösung eines Problems sehr individuell auf einer Webseite bewegen (Chi et al., 2003), was Webseiten aufgrund der unterschiedlichen Lösungsansätze ebenfalls schwer evaluierbar macht (Spool & Schroeder, 2001). Mithilfe von HolidayCheck.de sollte dementsprechend die wissenschaftliche Güte von IntuiBeat-F anhand der wissenschaftlichen Gütekriterien *Gründlichkeit* und *Gültigkeit* im Vergleich zum vierten Experiment nicht nur unter Berücksichtigung einer anderen Domäne, sondern auch unter Berücksichtigung einer stärker intuitiv benutzbaren Softwareanwendung

(vergleiche IPO.Rack als weniger intuitiv benutzbare Softwareanwendung mit vielen Nutzungsproblemen) und damit auch bei einer Anwendung mit höchstwahrscheinlich weniger Nutzungsproblemen nachgewiesen werden.

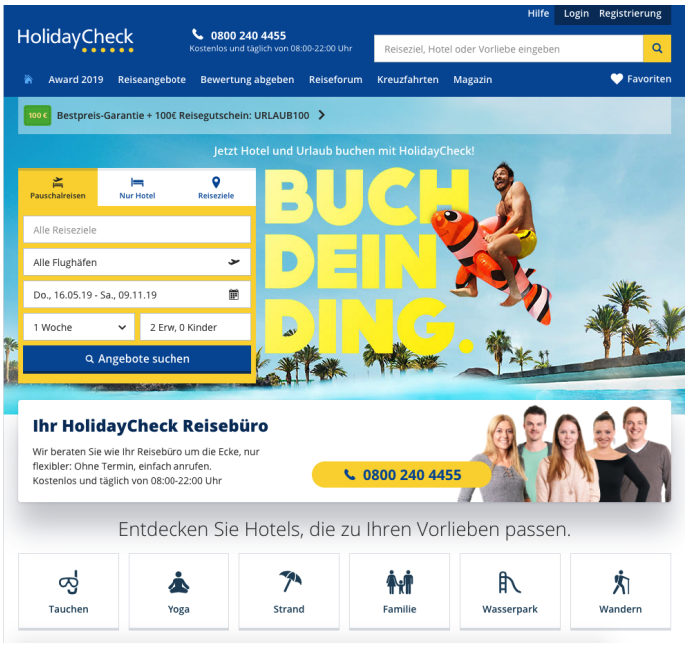


Abbildung 6.7. Die im Rahmen des fünften Experiments als Untersuchungsgegenstand verwendete Hotelbuchungswebseite *HolidayCheck.de* der HolidayCheck AG (2017).

Es wurden, auf Basis eigener Nutzungserfahrungen mit Reiseportalen, sechs experimentelle Aufgaben gewählt (siehe Anhang B.5.1), die den Funktionsumfang von HolidayCheck.de so gut wie möglich repräsentieren sollten. Sie wurden von den Versuchspersonen in einer festen Reihenfolge bearbeitet. Alle Aufgaben wurden dabei in ein fiktives Reiseszenario eingebettet, in dem es um die Reise zweier Freunde nach Japan ging. Hierbei mussten die Versuchspersonen als Erstes nach einem bestimmten Zimmer in Tokio für einen bestimmten Reisezeitraum suchen, welches eine Reihe von angegebenen Anforderungen erfüllen musste (erste Aufgabe). Um dieses Zimmer für eine spätere Buchung im Auge zu behalten, mussten die Versuchspersonen dieses Zimmer im Anschluss auf ihre Beobachtungsliste setzen (zweite Aufgabe). Daraufhin mussten die Versuchspersonen das beobachtete Zimmer buchen (dritte Aufgabe). Anschließend mussten die Versuchspersonen ihren fiktiven Aufenthalt in der Unterkunft über das Reiseportal bewerten (vierte Aufgabe). Um bei ihrer bevorstehenden Reise mit anderen Reisenden in Kontakt treten zu können, mussten sich die Versuchspersonen als nächstes mit einem bereitgestellten Benutzerkonto bei HolidayCheck.de anmelden und dort ihr Benutzerprofil mit bestimmten Informationen (z.B. Lieblingsreiseziele) ergänzen (fünfte Aufgabe). Abschließend mussten die Versuchspersonen über das Forum des Reiseportals nach Gleichgesinnten suchen, die sich während des gebuchten Reisezeitraums ebenfalls in Japan aufhielten und diese nach einem Treffen in Osaka fragen (sechste Aufgabe). Bei der Vorbereitung der Aufgaben wurden drei vorgefertigte Profile verwendet, um zu verhindern, dass nicht wirklich andere Personen kontaktiert werden. Darüber hinaus wurde bei den vorgelegten Aufgabenbeschreibungen darauf geachtet, dass die Buchungs- und Bewertungsaufgabe nicht wirklich von den Versuchspersonen

bis zum Ende durchgeführt wurden, sondern vor der eigentlichen Buchung bzw. Bewertung endete.

Vorerfahrung bei der Nutzung von Hotelbuchungswebseiten

Die Vorerfahrung der Versuchspersonen bezüglich der Hotelbuchungswebseite *HolidayCheck.de* wurde mit einem hierfür konstruierten TFQ als Kontrollvariable erhoben (KV_ Vorerfahrung), welcher wie in den Experimenten zuvor, in Papierform administriert wurde. Als Items für diesen Fragebogen wurden lediglich „HolidayCheck.de“ selbst und „Andere Hotelbuchungswebsite?“ genutzt, wobei letztere eine Wildcard darstellte und die Angabe einer beliebigen bekannten Hotelbuchungswebsite erlaubte. Es wurde bei der Konstruktion des TFQ im Vergleich zu den TFQs aus einigen vorigen Experimenten auf die Nennung konkreter Softwareanwendungen als Beispiele verzichtet, da es heute ein Überangebot an Hotelbuchungswebseiten und Reiseportalen gibt, was eine repräsentative Auswahl erschwert. Da wahrscheinlich jeder der als Stichprobe fungierenden Studierenden der Medienkommunikation und Mensch-Computer-Systeme schon einmal eine Reise oder ein Hotel über eine derartige Plattform gebucht hat, erschien die Wildcard ausreichend, um die damit verbundene Expertise erfassen zu können.

Wie bei den anderen konstruierten TFQs, beurteilten die Versuchspersonen diese beiden Items dann bezüglich deren Nutzungshäufigkeit (Wertebereich: 0 - 6) und des genutzten Funktionsumfangs (Wertebereich: 0 - 4). Im Anschluss wurde dann für jede Dimension (d.h. Nutzungshäufigkeit und Funktionsumfang) ein Mittelwert gebildet und daraufhin ein Gesamtmittelwert über beide Mittelwerte als Operationalisierung der *Vorerfahrung bei der Nutzung von Hotelbuchungswebseiten* berechnet (Wertebereich: 0 - 5). Die Entscheidung für eine Mittelwertbildung anstelle einer Summenberechnung wurde im Rahmen des ersten Experiments ausführlich begründet, weswegen an dieser Stelle lediglich darauf verwiesen wird (siehe Absatz „Vorerfahrung bei der Nutzung von CUIs“ des Teilabschnitts 5.1.4.3). Da kein einheitlicher TFQ existiert und dieser jeweils für die getesteten Softwareanwendungen erstellt werden musste, ist der in diesem Experiment genutzte TFQ vollständig in Anhang B.5.2 dieser Arbeit zu finden.

6.2.3.4 Versuchsdurchführung

Die Versuchsdurchführung des fünften Experiments war im Vergleich zu der Versuchsdurchführung des vierten Experiments mit der Ausnahme identisch, dass von den Versuchspersonen mit der Hotelbuchungswebseite *HolidayCheck.de* anstelle von der Regalplanungssoftware *IPO.Rack* gearbeitet wurde, weswegen an dieser Stelle für die Beschreibung der Versuchsdurchführung auf den entsprechenden Teilabschnitt 6.1.3.4 des vierten Experiments verwiesen werden soll.

6.2.3.5 Statistische Auswertung

Die statistische Auswertung des fünften Experiments war identisch mit der statistischen Auswertung des vierten Experiments, weswegen für genauere Informationen auf den ent-

sprechenden Teilabschnitt 6.1.3.5 des vierten Experiments verwiesen werden soll. Da die *Vorerfahrung bei der Nutzung von Hotelbuchungswebseiten*, die als Kontrollvariable erhoben wurde, einen ungewollten Einfluss auf die Gründlichkeit und Gültigkeit der verglichenen Methoden haben könnte, musste eine derartige Konfundierung vor der Überprüfung der Gründlichkeit und Gültigkeit von IntuiBeat-F ausgeschlossen werden (siehe Teilabschnitt 6.2.4.2). Hierzu wurden *t*-Tests für unabhängige Stichproben bezüglich der Vorerfahrung, sowie Pearson-Produkt-Moment-Korrelationen zwischen der Vorerfahrung und den abhängigen Variablen innerhalb der beiden Ausprägungen der unabhängigen Variable gerechnet (siehe Teilabschnitt 6.6). Da in beiden Fällen keine signifikanten Ergebnisse festgestellt werden konnten (d.h. kein ungewollter Einfluss der Vorerfahrung auf Gründlichkeit und Gültigkeit), wurde von einer univariaten Kovarianzanalyse (ANCOVA) abgesehen (Döring & Bortz, 2016; Field, 2017) und sich stattdessen für eine Datenauswertung mithilfe von *t*-Tests für unabhängige Stichproben ohne Berücksichtigung der Vorerfahrung bei der Nutzung von Hotelbuchungswebseiten als Kovariate entschieden (siehe Abschnitt 6.2.4). Die Überprüfung der statistischen Voraussetzungen der Pearson-Produkt-Moment-Korrelationen und der *t*-Tests werden im folgenden Ergebnisteil berichtet (siehe Teilabschnitt 6.2.4.1).

6.2.4 Ergebnisse

Im folgenden Abschnitt werden die Ergebnisse bezüglich der in den Teilabschnitten 6.2.1 und 6.2.2 beschriebenen Hypothesen deskriptiv und inferenzstatistisch berichtet. Vor der eigentlichen Datenanalyse wird zunächst auf die Überprüfung der statistischen Voraussetzungen eingegangen. Dabei wurde bei allen abhängigen Variablen und Kontrollvariablen ein metrisches Skalenniveau angenommen.

6.2.4.1 Überprüfung der statistischen Voraussetzungen

Überprüfung von Ausreißern

Das Vorgehen bei der Ausreißeranalyse des fünften Experiments war identisch mit dem Vorgehen des vierten Experiments, weswegen für genauere Informationen auf den entsprechenden Teilabschnitt 6.1.4.1 des vierten Experiments verwiesen werden soll. Es mussten auf diese Weise bei keiner der analysierten abhängigen Variablen und Kontrollvariablen Werte ausgeschlossen werden.

Überprüfung der Voraussetzung der Normalverteilung

Die univariate Normalverteilung der abhängigen Variablen und Kontrollvariablen wurde für jede Ausprägung der unabhängigen Variable mittels Kolmogorov-Smirnov-Tests ($p \geq .05$, siehe Field, 2017) und Sichtprüfung anhand eines Q-Q-Diagramms geprüft. Dabei konnten auf Basis dieser beiden Kriterien bei AV_Gründlichkeit_WenigerStrikt im Rahmen der Überprüfung der Gründlichkeit (Hypothese H1), bei AV_Gültigkeit_Strikt im Rahmen der Überprüfung der Gültigkeit (Hypothese H2) und bei der AV_ZeitlicheAnwendungseffizienz im Rahmen der Überprüfung der zeitlichen Anwendungseffizienz (Hypothese

H3) keine Normalverteilung in beiden Gruppen festgestellt werden. Aufgrund der Tatsache, dass ein ungepaarter *t*-Test bei etwa gleich großen Gruppen robust gegenüber Verletzungen der Normalverteilungsannahme ist (Glass et al., 1972; Pagano, 2006; Salkind, 2010; Wilcox, 2011), wurde sich für eine eindeutige Interpretation und Vergleichbarkeit der Ergebnisse (d.h. Mittelwerte statt Medianen) der verschiedenen Experimente gegen Transformationen und entsprechende nonparametrische Verfahren zur Überprüfung der Hypothesen (d.h. H1, H2 und H3) entschieden.

Überprüfung der Voraussetzung der Homoskedastizität

Zur Überprüfung der Homoskedastizität zwischen den Ausprägungen der unabhängigen Variable kamen Levene-Tests zum Einsatz (siehe Field, 2017), welche im Rahmen der Überprüfung der Gründlichkeit (Hypothese H1), der Überprüfung der Gültigkeit (Hypothese H2) und der Überprüfung der zeitlichen Anwendungseffizienz (Hypothese H3) gerechnet und bei allen abhängigen Variablen und Kontrollvariablen ($p \geq .05$, siehe Field, 2017) außer bei AV_Gründlichkeit_WenigerStrikt Varianzhomogenität bestätigen konnten. Die Freiheitsgrade für den dazugehörigen *t*-Test im Rahmen der Überprüfung der Gründlichkeit (Hypothese H1) wurden dementsprechend korrigiert.

6.2.4.2 Überprüfung einer möglichen Konfundierung durch Vorerfahrung bei der Nutzung von Hotelbuchungswebseiten

Die Vorerfahrung bei der Nutzung von Hotelbuchungswebseiten (Wertebereich: 0 - 5) unterschied sich nicht signifikant zwischen den beiden Ausprägungen der unabhängigen Variable (Nutzertest mit retrospektivem Think-Aloud-Protokoll: $M = 1.44$; $SD = .78$; IntuiBeat-F: $M = .98$, $SD = .74$), $t(22) = -1.48$, $p = .154$, d = .61. Laut J. Cohen (1988) kann dieser Effekt als mittel interpretiert werden ($d \leq .8$). Eine konservative post-hoc Analyse der Teststärke bezüglich KV_Vorerfahrung mithilfe von G*Power (Faul et al., 2009) mit $df = 22$ und einer angenommenen mittleren Effektstärke (d = .61) ergab lediglich eine geringe Teststärke (1 - $\beta = .29$) im Zuge der Auswertung der Vorerfahrung bei der Nutzung von Hotelbuchungswebseiten. Um jedoch eine ausreichend große Power (1 - $\beta \geq .80$) beim festgestellten Effekt erzielen zu können, wären pro Gruppe 44 Versuchspersonen nötig gewesen, was aufgrund des straffen Zeitplans im Anwenderprojekt, des dortigen Fokus auf qualitative Ergebnisse, des Verständnisses der Meta-Evaluation von IntuiBeat-F als Nebenprodukt und der personellen Einschränkungen im Projekt 3D-GUIde nicht möglich gewesen wäre. Auch bei Annahme einer großen Effektstärke (d = .8) hätten immerhin noch 26 Versuchspersonen pro Gruppe getestet werden müssen, was im Hinblick auf die Einschränkungen durch das Anwenderprojekt nicht notwendig erschien.

Da die erhobene Vorerfahrung bei der Nutzung von Hotelbuchungswebseiten dennoch einen ungewollten Einfluss auf den Vergleich der beiden Ausprägungen der unabhängigen Variable haben könnte, wurde für die Kontrollvariable mithilfe von Pearson-Produkt-Moment-Korrelationen sichergestellt, dass kein Zusammenhang zwischen dieser und den abhängigen Variablen innerhalb der beiden Ausprägungen der unabhängigen Variable besteht ($p > .05$). Die Kontrollvariable KV_Vorerfahrung wurde dementsprechend nicht als Kovariate

in der statistischen Auswertung berücksichtigt, da kein linearer Zusammenhang zwischen dieser und den abhängigen Variablen bestand (siehe Döring & Bortz, 2016; Field, 2017).

6.2.4.3 Überprüfung der Gründlichkeit

Wie erwartet, lag die Gründlichkeit (%) von IntuiBeat-F (M = .29, SD = .15) signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll (M = .19, SD = .05), wenn alle mit IntuiBeat-F abgeleiteten Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) berücksichtigt wurden (H1.A), $t(14.05)$ = - 2.18, p = .047, d = .89. Es war laut J. Cohen (1988) ein großer Effekt feststellbar (d $\geq$ .8). Entgegen der Erwartungen lag die Gründlichkeit beim Nutzertest mit retrospektivem Think-Aloud-Protokoll (M = .20, SD = .06) signifikant höher als bei IntuiBeat-F (M = .14, SD = .06), wenn nur die auf Basis des Algorithmus von IntuiBeat-F abgeleiteten Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F) berücksichtigt wurden (H1.B), $t(22)$ = 2.81, p = .010, d = 1. Es war laut J. Cohen (1988) ein großer Effekt feststellbar (d $\geq$ .8).

6.2.4.4 Überprüfung der Gültigkeit

Wie erwartet, lag die Gültigkeit (%) von IntuiBeat-F (M = .79, SD = .10) signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll (M = .64, SD = .14), wenn alle mit IntuiBeat-F abgeleiteten Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) berücksichtigt wurden (H2.A), $t(22)$ = - 2.94, p = .008, d = 1.23. Es war laut J. Cohen (1988) ein großer Effekt feststellbar (d $\geq$ .8). Auch wenn ausschließlich die auf Basis des Algorithmus von IntuiBeat-F abgeleiteten Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F) berücksichtigt werden (H2.B), lag die Gültigkeit von IntuiBeat-F (M = .97, SD = .08), wie erwartet, signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll (M = .64, SD = .14), $t(22)$ = - 6.97, p = .000, d = 2.89. Es war laut J. Cohen (1988) hier ebenfalls ein großer Effekt feststellbar (d $\geq$ .8).

6.2.4.5 Überprüfung der zeitlichen Anwendungseffizienz

Wie erwartet, lag die zeitliche Anwendungseffizienz (s) von IntuiBeat-F in Form einer geringeren durchschnittlichen Problemidentifikationszeit für reale Nutzungsprobleme (M = 268.59, SD = 102.38) signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll (M = 358.77, SD = 105.43), $t(22)$ = 2.13, p = .045, d = .86. Es war laut J. Cohen (1988) ein großer Effekt feststellbar (d $\geq$ .8).

6.2.4.6 Explorative Datenanalyse

Wie bereits in Teilabschnitt 6.1.3.3 des vierten Experiments angesprochen, wurden zusätzlich eine Reihe von Metriken für eine explorative Datenanalyse berechnet, mit deren Hilfe verstanden werden kann, warum es zu den Unterschieden bezüglich Gründlichkeit (siehe Teilabschnitt 6.2.4.3) und Gültigkeit (siehe Teilabschnitt 6.2.4.4) bei der strikten und weniger strikten Anwendung von IntuiBeat-F gekommen ist. Im Folgenden sollen diese Metriken nun berichtet werden.

Tabelle 6.4. *Metriken zur explorativen Datenanalyse von IntuiBeat-F und des Nutzertests mit retrospektivem Think-Aloud-Protokoll im Rahmen des fünften Experiments (NPs: Nutzungsprobleme; KEs: kritische Ereignisse; RP-KEs: mit Rhythmus-Peaks assoziierte kritische Ereignisse; Alle NPs aus KEs ≈ weniger strikte Anwendung von IntuiBeat-F; Nur NPs aus RP-KEs ≈ strikte Anwendung von IntuiBeat-F).*

Metrik	Experiment 5	
	IntuiBeat-F	Nutzertest
Trefferrate NPs (%):		
- Alle NPs aus KEs	79.91	76.56
- Nur NPs aus RP-KEs	100	-
Trefferrate reale NPs (%):		
- Alle NPs aus KEs	77.68	75.52
- Nur NPs aus RP-KEs	97.40	-
Fehlalarmrate NPs (%):		
- Alle NPs aus KEs	20.09	23.44
- Nur NPs aus RP-KEs	0	-
Fehlalarmrate reale NPs (%):		
- Alle NPs aus KEs	22.32	24.48
- Nur NPs aus RP-KEs	2.60	-
Anzahl NPs:		
- Alle NPs aus KEs	50	37
- Nur NPs aus RP-KEs	37	-
Anzahl realer NPs:		
- Alle NPs aus KEs	47	36
- Nur NPs aus KEs mit RPs	36	-
Anteil realer NPs an NPs (%):		
- Alle NPs aus KEs	94	97.30
- Nur NPs aus RP-KEs	97.30	-

Explorative Datenanalyse von IntuiBeat-F (weniger strikte Anwendung)

Mithilfe von IntuiBeat-F konnten über alle Versuchspersonen hinweg 224 kritische Ereignisse erkannt werden, wovon unter Berücksichtigung der in der Abbildung 6.2 in Teilabschnitt 6.1.3 beschriebenen Arbeitsschritte zur Ableitung von Nutzungsproblemen (Burmester, 2016) und der handlungsorientierten Fehlertaxonomie (Zapf et al., 1989) 179 kritische Ereignisse (79.91 %) zu insgesamt 50 einzigartigen Nutzungsproblemen konsolidiert

wurden (d.h. weniger strikte Anwendung von IntuiBeat-F). Von diesen 50 Nutzungsproblemen, die sich aus fünf der 224 kritischen Ereignisse (2.23 %) ableiteten, wurde bei drei Nutzungsproblemen ein Bewegungsfehler (6 %) als Ursache klassifiziert. Es handelt sich dementsprechend nur bei 47 der 50 Nutzungsprobleme (94 %) auch um reale Nutzungsprobleme, die zur Sicherstellung einer intuitiven Benutzung beseitigt werden müssen. Diese realen Nutzungsprobleme wurden aus 174 von 224 kritischen (77.68 %) abgeleitet. Die restlichen 45 der 224 kritischen Ereignisse (20.09 %) stellten lediglich (zeitliche) Ineffizienzen im Sinne der handlungsorientierten Fehlertaxonomie dar und bildeten dadurch ebenfalls keine realen Nutzungsprobleme ab. Bei der Einteilung der Nutzungsprobleme in die Fehlerkategorien der handlungsorientierten Fehlertaxonomie von Zapf et al. (1989) konnte ein Cohens κ von .80 festgestellt werden, was laut Landis und Koch (1977) als eine beachtliche Übereinstimmung interpretiert werden kann und damit auf einem ähnlich hohen Niveau wie im Vorgängerexperiment lag ($\kappa = .78$).

Zusammenfassend kann festgehalten werden, dass die Trefferrate von IntuiBeat-F bei wenig strikter Anwendung bezogen auf alle Nutzungsprobleme damit bei 79.91 % und die Fehlalarmrate bei 20.09 % lag. Betrachtet man nur die aus kritischen Ereignissen abgeleiteten realen Nutzungsprobleme, reduziert sich die Trefferrate bei wenig strikter Anwendung entsprechend auf 77.68 % und die Fehlalarmrate steigt auf 22.32 % (siehe Tabelle 6.4).

Explorative Datenanalyse von IntuiBeat-F (strikte Anwendung)

Betrachtet man ausschließlich die mithilfe von Rhythmus-Peaks erkannten kritischen Ereignisse (d.h. strikte Anwendung von IntuiBeat-F), wurden insgesamt 113 Rhythmus-Peaks generiert, wo sich entsprechend 77 Rhythmus-Peaks (68.14 %) auf 37 Nutzungsprobleme bzw. 75 Rhythmus-Peaks (66.37 %) auf 36 reale Nutzungsprobleme verteilten. Zwei der 114 Rhythmus-Peaks (1.75 %) wiesen damit auf einen Bewegungsfehler hin. Die restlichen 36 der 113 Rhythmus-Peaks (31.86 %) gingen dementsprechend nicht in Nutzungsprobleme ein und führten nur in einem Fall zur Entdeckung einer (zeitlichen) Ineffizienz im Sinne der handlungsorientierten Fehlertaxonomie (0.89 %). Betrachtet man ausschließlich die durch jeweils einen Rhythmus-Peak erkannten 77 der 224 (34.38 %) insgesamt mit IntuiBeat-F gefundenen kritischen Ereignisse, gingen hierbei alle mithilfe von Rhythmus-Peaks identifizierten kritischen Ereignisse (100 %) in insgesamt 37 der insgesamt mit IntuiBeat-F gefundenen 50 Nutzungsprobleme (74 %) ein. Beschränkt man sich nur auf die mit IntuiBeat-F abgeleiteten realen Nutzungsprobleme, konnten 75 der 77 durch Rhythmus-Peaks identifizierten, auf reale Nutzungsprobleme hinweisenden, kritischen Ereignisse (97.40 %) 36 der 37 mithilfe von IntuiBeat-F abgeleiteten Nutzungsprobleme (97.30 %) identifizieren. Schließlich gingen zwei (2.60 %) von insgesamt 77 mithilfe von Rhythmus-Peaks identifizierten kritische Ereignisse in ein Nutzungsproblem ein, welches einen Bewegungsfehler als Ursache hatte. Dementsprechend lag die Trefferrate bei strikter Anwendung von IntuiBeat-F bei der Ableitung aller Nutzungsprobleme bei 100 % und die Fehlalarmrate bei 0 %, was einer Verbesserung von rund 20 % gegenüber der wenig strikten Anwendung entsprach. Betrachtet man nur die durch die strikte Anwendung von IntuiBeat-F abgeleiteten realen Nutzungsprobleme kann eine Trefferrate von 97.40 % und eine Fehlalarmrate von 2.6 % festgestellt werden, was ebenfalls einer Verbesserung von 20 % gegenüber der wenig strikten Anwendung entsprach (siehe Tabelle 6.4).

Zusammenfassend kann also festgehalten werden, dass vom Evaluator lediglich 68.14 % aller 113 Rhythmus-Peaks für die Ableitung von Nutzungsproblemen berücksichtigt und damit 31.86 % der Rhythmus-Peaks ignoriert wurden. Betrachtet man lediglich reale Nutzungsprobleme, wurden entsprechend nur 66.37 % aller 113 Rhythmus-Peaks für die Ableitung von realen Nutzungsproblemen genutzt und damit sogar 33.63 % der Rhythmus-Peaks vernachlässigt. In diesem Zusammenhang wurden lediglich 34.38 % von den gesamten 224 kritischen Ereignissen mithilfe der 77 Rhythmus-Peaks entdeckt, was bezogen auf die 179 auf Nutzungsprobleme hinweisenden kritischen Ereignisse 42.02 % (77 Rhythmus-Peaks) und bezogen auf die 174, auf reale Nutzungsprobleme hinweisenden, kritischen Ereignisse 43.10 % (75 Rhythmus-Peaks) darstellte. Dementsprechend wurden 65.62 % aller kritischen Ereignisse unabhängig von Rhythmus-Peaks protokolliert. Bezogen auf alle Nutzungsprobleme wurden damit 56.98 % und bezogen auf alle realen Nutzungsprobleme 56.90 % der kritischen Ereignisse unabhängig festgehalten. Aus diesem Grund konnten bei der ausschließlichen Berücksichtigung von Rhythmus-Peaks (d.h. strikte Anwendung von IntuiBeat-F) bei IntuiBeat-F 13 Nutzungsprobleme (d.h. nur 74 % noch auffindbar) und 11 reale Nutzungsprobleme weniger (d.h. nur 76.60 % noch auffindbar) bei gleichzeitiger Steigerung der Trefferrate bzw. Senkung der Fehlalarmrate beobachtet werden (siehe Tabelle 6.4).

Explorative Datenanalyse des Nutzertests mit retrospektivem Think-Aloud-Protokoll

Mithilfe des Nutzertests mit retrospektivem Think-Aloud-Protokoll konnten insgesamt 192 kritische Ereignisse erkannt werden, wovon unter Berücksichtigung der in der Abbildung 6.2 in Teilabschnitt 6.1.3 beschriebenen Arbeitsschritte zur Ableitung von Nutzungsproblemen (Burmester, 2016) und der handlungsorientierten Fehlertaxonomie (Zapf et al., 1989) 147 kritische Ereignisse (76.56 %) zu insgesamt 37 einzigartigen Nutzungsproblemen konsolidiert wurden, wovon zwei kritische Ereignisse (1.04 %) denselben Bewegungsfehler als Ursache hatten (5.41 %) und es sich damit nur bei 36 der 37 Nutzungsprobleme (97.30 %) um reale Nutzungsprobleme handelte, die damit aus 145 der 192 kritischen Ereignisse abgeleitet wurden (75.52 %). Die restlichen 45 der 192 kritischen Ereignisse (23.44 %) stellten lediglich (zeitliche) Ineffizienzen im Sinne der handlungsorientierten Fehlertaxonomie und damit auch keine echten Nutzungsprobleme im Bezug auf intuitive Benutzung dar. Dementsprechend ließen sich 36 reale Nutzungsprobleme aus 145 der 192 kritischen Ereignisse ableiten (72.52 %).

Zusammenfassend kann festgehalten werden, dass die Trefferrate bezogen auf alle aus kritischen Ereignissen abgeleiteten Nutzungsprobleme geringer als bei IntuiBeat-F bei 76.56 % und die Fehlalarmrate höher als bei IntuiBeat-F bei 23.44 % lag. Betrachtet man nur die aus kritischen Ereignissen abgeleiteten realen Nutzungsprobleme, reduziert sich die Trefferrate entsprechend auf 75.52 % und die Fehlalarmrate steigt auf 24.48 % (siehe Tabelle 6.4).

Qualitative Ursachenanalyse mit beiden Methoden gefundener Nutzungsprobleme bei wenig strikter Anwendung von IntuiBeat-F

Beurteilt man die mithilfe von IntuiBeat-F (d.h. weniger strikte Anwendung) und dem Nutzertest mit retrospektivem Think-Aloud-Protokoll abgeleiteten Nutzungsprobleme mit der handlungsorientierten Fehlertaxonomie qualitativ, stellt man bei Berücksichtigung aller kritischen Ereignisse folgende Einteilung fest (siehe Abbildung 6.8). Im Falle von IntuiBeat-F konnten zwei Urteilsfehler (15.39 %), zwei Gewohnheitsfehler (15.39 %), fünf Erkennensfehler (38.46 %), ein Merk-/Vergessensfehler (7.69 %), zwei Denkfehler (15.38 %) und ein Bewegungsfehler (7.69 %) als Ursachen für die 13 Nutzungsprobleme klassifiziert werden. Damit überwogen Ursachen, die der perzeptiv-begrifflichen Ebene zugeschrieben werden können. Bezüglich des Nutzertests mit retrospektivem Think-Aloud-Protokoll konnten ein Urteilsfehler (100 %) als Ursache für das eine einzigartige Nutzungsproblem klassifiziert werden. Die Ursache lag damit allein auf der intellektuellen Regulationsebene. Die Ursachen für die von beiden Methoden unter Berücksichtigung aller kritischen Ereignisse abgeleiteten 37 Nutzungsprobleme waren neun Gewohnheitsfehler (24.32 %), fünf Urteilsfehler (13.51 %), zwei Merk-/Vergessensfehler (5.41 %), drei Unterlassensfehler (8.11 %), 13 Erkennensfehler (35.13 %), drei Denkfehler (8.11 %) und zwei Bewegungsfehler (5.41 %). Die von beiden Methoden gefundenen Ursachen teilen sich dementsprechend mit 27.02 % auf die intellektuelle, mit 67.57 % auf die perzeptiv-begriffliche und mit 5.41 % auf die sensomotorische Regulationsebene auf.

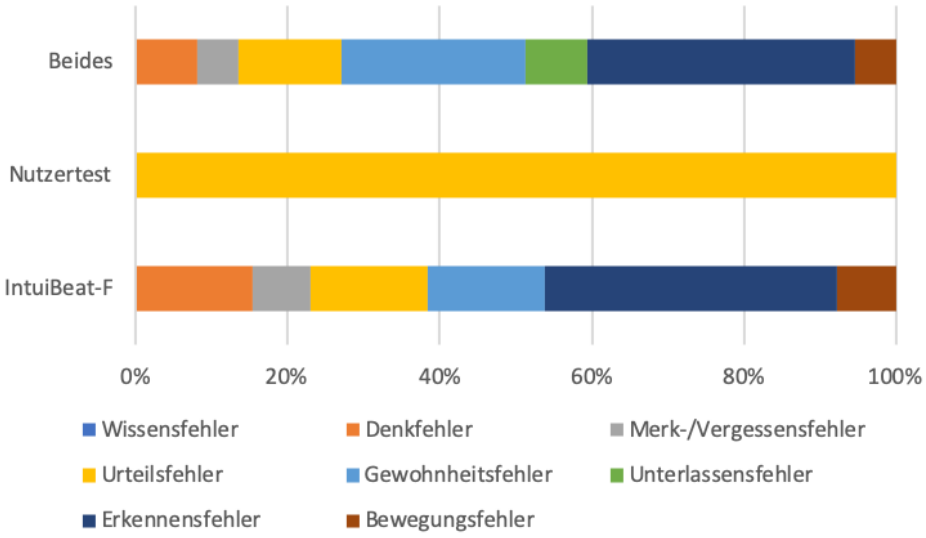


Abbildung 6.8. Ursachenklassifikation mithilfe der handlungsorientierten Fehlertaxonomie (Zapf, Brodbeck, & Prümper, 1989) der mit IntuiBeat-F (weniger strikte Anwendung) und einem Nutzertest mit retrospektivem Think-Aloud-Protokoll abgeleiteten Nutzungsprobleme im Rahmen des fünften Experiments.

Qualitative Ursachenanalyse mit beiden Methoden gefundener Nutzungsprobleme bei strikter Anwendung von IntuiBeat-F

Berücksichtigt man bei dieser qualitativen Beurteilung bei IntuiBeat-F ausschließlich die mithilfe des Algorithmus abgeleiteten Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F), ergibt sich folgende geänderte Einteilung (siehe Abbildung 6.9). Im Falle von IntuiBeat-F konnten ein Gewohnheitsfehler ($11.\overline{11}$ %), zwei Urteilsfehler ($22.\overline{22}$ %), ein Denkfehler ($11.\overline{11}$ %), vier Erkennensfehler ($44.\overline{44}$ %) und ein Merk-/Vergessensfehler ($11.\overline{11}$ %) als Ursachen für die neun Nutzungsprobleme klassifiziert werden, wodurch die Mehrheit der Ursachen der perzeptiv-begrifflichen Ebene zugeordnet werden kann. Bezüglich des Nutzertests mit retrospektivem Think-Aloud-Protokoll konnten drei Urteilsfehler (27.27 %), sechs Erkennensfehler (54.55 %) und zwei Bewegungsfehler (18.18 %) als Ursachen für die 11 Nutzungsfehler identifiziert werden, wodurch sich die zugrunde liegenden Ursachen hauptsächlich der perzeptiv-begrifflichen Regulationsebene zuordnen lassen. Die Ursachen für die, von beiden Methoden unter Berücksichtigung nur der bei IntuiBeat-F algorithmisch abgeleiteten, 28 Nutzungsprobleme waren neun Gewohnheitsfehler (32.14 %), drei Urteilsfehler (10.71 %), drei Unterlassensfehler (10.71 %), sieben Erkennensfehler (25.00 %), zwei Merk-/Vergessensfehler (7.14 %), drei Denkfehler (10.71 %) und ein Bewegungsfehler (3.57 %). Die von beiden Methoden gefundenen Ursachen teilen sich dementsprechend mit 28.57 % auf die intellektuelle, mit 67.86 % auf die perzeptiv-begriffliche und mit 3.57 % auf die sensomotorische Regulationsebene auf.

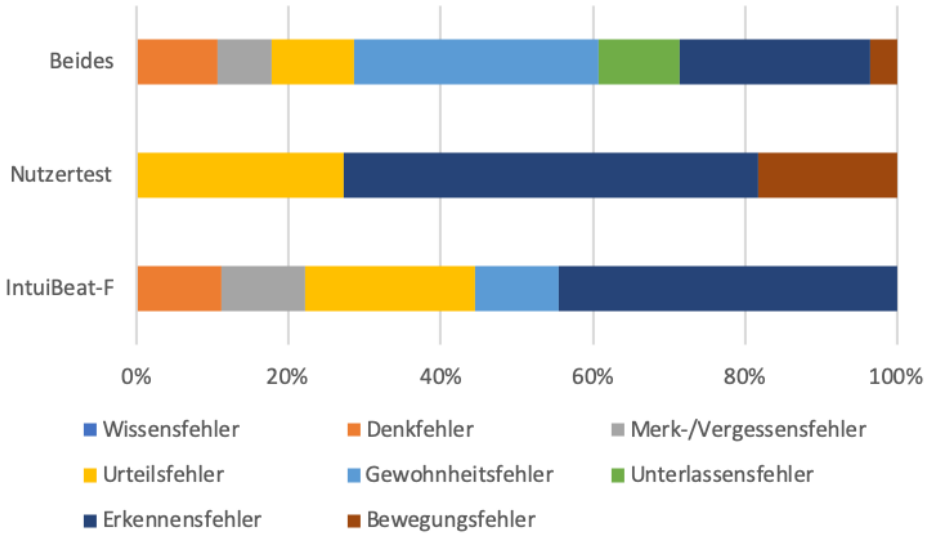


Abbildung 6.9. Ursachenklassifikation mithilfe der handlungsorientierten Fehlertaxonomie (Zapf, Brodbeck, & Prümper, 1989) der mit IntuiBeat-F (strikte Anwendung) und einem Nutzertest mit retrospektivem Think-Aloud-Protokoll abgeleiteten Nutzungsprobleme im Rahmen des fünften Experiments.

6.2.5 Diskussion

Im vorliegenden fünften Experiment wurde die wissenschaftliche Güte von IntuiBeat-F bezüglich der Gütekriterien *Gründlichkeit* und *Gültigkeit*, sowie der *zeitlichen Anwendungseffizienz*, als für das Projekt 3D-GUIde wichtiges Kriterium der praktischen Güte, beim Vergleich mit einem als Quasi-Außenkriterium fungierenden Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der Nutzung der Hotelbuchungswebseite *HolidayCheck.de* überprüft, bei der es sich auf Basis von Experteneinschätzungen und Nutzereinschätzungen (d.h. anhand eines TFQ) um eine eher stärker intuitiv benutzbare Software handelte, bei der es entsprechend zu wenigen, intuitive Benutzung beeinträchtigenden Nutzungsproblemen (d.h. reale Nutzungsprobleme) kommen sollte.

Anhand der statistischen Tests wurde zunächst überprüft, ob sich ein Unterschied bezüglich der Gründlichkeit zu Gunsten von IntuiBeat-F feststellen lässt (Hypothese H1). Anschließend wurde geprüft, ob auch bezüglich der Gültigkeit ein Unterschied zu Gunsten von IntuiBeat-F erkennbar ist (Hypothese H2). Abschließend wurde untersucht, inwiefern IntuiBeat-F eine höhere zeitliche Anwendungseffizienz als der Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der formativen Evaluation intuitiver Benutzung der Hotelbuchungswebseite *HolidayCheck.de* aufweisen kann (Hypothese H3). Im folgenden Verlauf werden die Hypothesen bezüglich der Gründlichkeit, der Gültigkeit und der zeitlichen Anwendungseffizienz unter Berücksichtigung der festgestellten Ergebnisse diskutiert, nachdem im Vorfeld auf eine mögliche Konfundierung durch die *Vorerfahrung bei der Nutzung von Hotelbuchungswebseiten* eingegangen wurde.

6.2.5.1 Überprüfung einer möglichen Konfundierung durch die Vorerfahrung bei der Nutzung von Hotelbuchungswebseiten

Die Vorerfahrung bei der Nutzung von Hotelbuchungswebseiten unterschied sich erwartungsgemäß nicht signifikant zwischen den beiden Ausprägungen der unabhängigen Variable *Art der formativen Evaluationsmethode*. Jedoch war die Teststärke aufgrund des mittleren beobachteten Effekts und der kleinen Stichprobe gering. Im Gegensatz zum vorigen Experiment (siehe Abschnitt 6.1), in dem hauptsächlich die verwendete Stichprobe alleine für die geringe Varianz in der Vorerfahrung potentiell verantwortlich gemacht wurde, könnte in diesem Experiment speziell die Operationalisierung der Vorerfahrung durch den TFQ selbst mit hoher Wahrscheinlichkeit in Zusammenhang mit der verwendeten Studierendenstichprobe verantwortlich für dieses Ergebnis sein. Wie in Abschnitt 3.6.1 beschrieben, erfasst der TFQ Vorerfahrung anhand der Dimensionen *Häufigkeit* und *Intensität*.

Beispielsweise stellt innerhalb des verwendeten TFQ die „tägliche“ Nutzung (entspricht „6“ auf der Likertskala) das obere Ende der in Zusammenhang mit der Häufigkeit verwendeten Likertskala dar, eine Ausprägung, die bei Hotelbuchungswebseiten im Kreis von Studierenden mit hoher Wahrscheinlichkeit nicht vorkommt. „Alle paar Monate“ (entspricht „2“ auf der Likertskala) oder „lediglich einmal oder zweimal verwendet“ (entspricht „1“ auf der Likertskala) entsprechen bei Hotelbuchungswebseiten unter Berücksichtigung der Studierendenstichprobe hingegen eher der Realität. Die Varianz ist bei der verwendeten Studierendenstichprobe dabei trotzdem als eher gering einzustufen, weswegen auch keine

großen Schwankungen in den Antworten zu erwarten sind. Auch bezüglich der Funktionsdimension, bei der „alle Features“ (entspricht „4“ auf der Likertskala) das obere Ende der in diesem Zusammenhang verwendeten Likertskala darstellt, ist keine hohe Varianz bei der verwendeten Stichprobe und dem genutzten Untersuchungsgegenstand zu erwarten, da Hotelbuchungswebseiten hauptsächlich für das Buchen von Hotels genutzt werden und andere Funktionen (z.B. integrierte Reiseforen) nicht im primären Fokus stehen. Dementsprechend ist mit hoher Wahrscheinlichkeit zu erwarten, dass sich die Antworten der Studierenden bei „genug Features, um damit arbeiten zu können“ einpendeln und hier ebenfalls nicht mit großen Schwankungen zu rechnen ist.

Die praktische Bedeutsamkeit der in diesem Zusammenhang ermittelten Effekte bzw. der Erkenntnisgewinn ist aufgrund der verwendeten Stichprobe und der mangelnden Sensitivität des TFQ entsprechend als eher unbedeutend einzustufen (Döring & Bortz, 2016), was sich auch in den deskriptiven Unterschieden erkennen lässt. Da darüber hinaus bei der Kontrollvariablen keine signifikanten Korrelationen bezüglich der Gründlichkeit, der Gültigkeit und der zeitlichen Anwendungseffizienz als wichtiger Teilaspekt praktischer Güte bei beiden formativen Evaluationsmethoden vorlagen ($p > .05$), lag mit hoher Wahrscheinlichkeit keine Konfundierung durch eine unterschiedliche Vorerfahrung bei der Nutzung von Hotelwebseiten vor.

6.2.5.2 Überprüfung der Gründlichkeit

Wie bereits beim vierten Experiment (siehe Abschnitt 6.1), zeigte sich auch im fünften Experiment unerwarteterweise die prognostizierte höhere Gründlichkeit von IntuiBeat-F gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der Evaluation einer stärker intuitiv benutzbaren Software lediglich inferenzstatistisch bei Berücksichtigung aller realen Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) und nicht bei ausschließlicher Berücksichtigung der durch Rhythmus-Peaks entdeckten realen Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F). Bei Berücksichtigung aller realen Nutzungsprobleme konnte IntuiBeat-F gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll eine signifikant höhere Gründlichkeit von 29 % im Vergleich zu 19 % (siehe Teilabschnitt 6.2.4.3) und damit ein nahezu ähnliches Verhältnis wie im vorangegangenen vierten Experiment (d.h. Gründlichkeit von IntuiBeat-F lag auch im Rahmen des vierten Experiments bei Berücksichtigung aller realen Nutzungsprobleme bei rund 30 %) zeigen (siehe Teilabschnitt 6.1.4.3). Wie die Ergebnisse des Vorgängerexperiments liegen die Ergebnisse des fünften Experiments bezüglich Gründlichkeit damit auch in einem ähnlichen Wertebereich von circa 20 bis 40 %, den Koutsabasis et al. (2007) allgemein für formative Evaluationsmethoden für Gebrauchstauglichkeit (d.h. nicht speziell intuitive Benutzung) angeben, wenn als Evaluatoren Studierende und damit Novizen eingesetzt werden, die zwar über eine große Kenntnis des Untersuchungsgegenstands verfügen, es ihnen jedoch noch noch an Erfahrung bei der Durchführung eines Nutzertests fehlt. Zusammenfassend lässt sich somit die in diesem Zusammenhang aufgestellte Hypothese H1.A auch bei der Evaluation einer stärker intuitiv benutzbaren Software bestätigen (siehe Tabelle 6.5).

Nichtsdestotrotz konnte, wie bereits beim Vorgängerexperiment (siehe Teilabschnitt 6.1.5), bei Berücksichtigung lediglich der auf einen Rhythmus-Peak rückführbaren realen Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F) kein signifikanter Unterschied zu Gunsten von IntuiBeat-F, und deskriptiv sogar eine Verschlechterung der Gründlichkeit um knapp 15 % gegenüber der weniger strikten Anwendung festgestellt werden (siehe Teilabschnitt 6.2.4.3). Betrachtet man hierzu die deskriptiven Daten des fünften Experiments unter diesem Aspekt genauer (siehe Teilabschnitt 6.2.4.6), ist dies nicht verwunderlich, da lediglich rund 70 % aller Rhythmus-Peaks für die Identifikation von kritischen Ereignissen, die auf auf reale Nutzungsprobleme hinweisen konnten, genutzt, nur rund 35 % dieser kritischen Ereignisse mithilfe von Rhythmus-Peaks entdeckt und demzufolge rund 23 % weniger reale Nutzungsprobleme mithilfe von IntuiBeat-F abgeleitet werden konnten. Eine Verschlechterung der Gründlichkeit von IntuiBeat-F um 15 % bei gleichzeitiger Verbesserung der Gründlichkeit beim Einsatz des Nutzertests mit retrospektivem Think-Aloud-Protokoll um 1 % ist die Folge (siehe Teilabschnitt 6.2.4.3), da der Nutzertest nun im Vergleich zu IntuiBeat-F einen größeren Anteil aller real verfügbaren Nutzungsprobleme identifiziert (d.h. es fallen hier keine realen Nutzungsprobleme weg). Da die Gesamtzahl aller vorhandenen realen Nutzungsprobleme durch eine Vereinigungsmenge der durch beide Evaluationsmethoden gefundenen Nutzungsprobleme bestimmt wurde, profitiert die Gründlichkeit des Nutzertests mit retrospektivem Think-Aloud-Protokoll von dieser Tatsache (siehe Teilabschnitt 6.1.3.3).

Damit befinden sich die Ergebnisse des fünften Experiments auf einem ähnlichen Niveau wie die Ergebnisse des Vorgängerexperiments (siehe Teilabschnitt 6.1.5), lediglich die Trefferrate bzw. Fehlalarmrate war im fünften Experiment bei beiden getesteten Methoden im Vergleich zum Vorgängerexperiment höher, da es sich bei einem größeren Anteil der identifizierten kritischen Ereignisse lediglich um (zeitliche) Ineffizienzen und damit nicht um Nutzungsprobleme im Sinne der handlungsorientierten Fehlertaxonomie (siehe Zapf et al., 1989) handelte. Wie bereits an gleicher Stelle im Rahmen der Diskussion des Vorgängerexperiments angesprochen (siehe Teilabschnitt 6.1.5), liegt eine mögliche Erklärung für die Verschlechterung der Gründlichkeit bei strikter Anwendung von IntuiBeat-F in der Tatsache, dass der Evaluator, statt die übrigen 30 % der Rhythmus-Peaks für die Identifikation von kritischen Ereignissen zu nutzen, einen großen Anteil an kritischen Ereignisse unabhängig von Rhythmus-Peaks protokolliert hat, aus denen sich letztendlich reale Nutzungsprobleme ableiten konnten. Dementsprechend ergibt sich die Frage, ob die knapp 65 % der unabhängig vom Algorithmus protokollierten kritischen Ereignisse, die auf reale Nutzungsprobleme hinweisen konnten, entdeckbar gewesen wären, wenn der Evaluator die restlichen knapp 30 % der Rhythmus-Peaks berücksichtigt hätte (siehe Teilabschnitt 6.2.4.6).

Wie bereits bei der entsprechenden Diskussion im Rahmen des vierten Experiments erwähnt (siehe Teilabschnitt 6.1.5), könnten die unberücksichtigten Rhythmus-Peaks echte Fehlalarme (d.h. hinter diesen verbirgt sich kein kritisches Ereignis) und nicht die Fehlinterpretation des Evaluators darstellen, falls er sie sogar überprüft hat. Da sowohl im gerade beschriebenen fünften Experiment als auch im Vorgängerexperiment über 30 % aller Rhythmus-Peaks bei der Datenauswertung nicht berücksichtigt wurden (d.h. sowohl bei einer wenig als auch einer stärker intuitiv benutzbaren Software) und damit im Vergleich zu einer vergleichbaren Arbeit von Akers et al. (2012), bei der ebenfalls automatisch generierte Indikatoren für kritische Ereignisse zur formativen Evaluation genutzt wurden,

keine große Variation zu erkennen ist (d.h. Evaluator ignorierte automatisch generierte Rhythmus-Peaks: 40 % im vierten und 30 % im fünften Experiment vs. Evaluator ignorierte automatisch generierte Backtracking-Operationen von Akers et al. (2012): 30 % bei Google SketchUp und 5 % bei Adobe Photoshop CS3), sollte ein weiteres Folgeexperiment durchgeführt werden.

In diesem Folgeexperiment soll durch Bereitstellung einer einfacheren Zuordnungsmöglichkeit für den Evaluator explizit ausgeschlossen werden, dass insbesondere die umständliche manuelle Zuordnung von Rhythmus-Peaks zu Videostellen während des retrospektiven Interviews für die geringe Berücksichtigung von Rhythmus-Peaks und die hohe Anzahl von unabhängig von Rhythmus-Peaks protokollierten kritischen Ereignissen, die auf reale Nutzungsprobleme hinweisen konnten, für die geringere Gründlichkeit bei strikter Anwendung von IntuiBeat-F verantwortlich war. Erst dann ließe sich die mangelnde Gründlichkeit von IntuiBeat-F bei der Evaluation einer stärker und einer weniger intuitiv benutzbarer Software allgemein auf die hohe Fehlalarmrate von IntuiBeat-F zurückführen. Als Untersuchungsgegenstand würde sich hierfür eine weniger intuitiv benutzbare Software eignen, da dort eine zusätzliche Unterstützungsmöglichkeit für die Zuordnung von Rhythmus-Peaks zu entsprechenden Videostellen besonders auf ihre Wirksamkeit getestet werden kann. In einer solchen Software sind im Gegensatz zu einer stärker intuitiv benutzbaren Software erwartungsgemäß mehr Nutzungsprobleme zu finden, die wiederum zu mehr Rhythmus-Peaks führen können. Der Evaluator muss demzufolge die Zuordnung in kurzer Zeit während des retrospektiven Interviews vornehmen, weswegen seine Gründlichkeit ohne zusätzliche Unterstützung darunter leiden würde.

Zusammenfassend kann IntuiBeat-F somit wie beim Vorgängerexperiment, wo eine weniger intuitiv benutzbare Software formativ evaluiert wurde, auch bei der Evaluation einer stärker intuitiv benutzbaren Software aufgrund der hohen Anzahl versäumter kritischer Ereignisse keine Gründlichkeit attestiert werden, wenn IntuiBeat-F strikt angewendet und dadurch lediglich Nutzungsprobleme bei der Bestimmung der Gründlichkeit, die auf Rhythmus-Peaks rückführbar sind, berücksichtigt werden. Aus diesem Grund muss auch die in diesen Zusammenhang aufgestellte Hypothese H1.B bei der Evaluation einer stärker intuitiv benutzbaren Software leider verworfen werden (siehe Tabelle 6.5). Aufgrund der Tatsache, dass, sowohl bei der Evaluation einer wenig intuitiv benutzbaren Software als auch bei der Evaluation einer stärker intuitiv benutzbaren Software, ein geringes Ausmaß an berücksichtigten Rhythmus-Peaks, eine hohes Ausmaß an unabhängig von Rhythmus-Peaks identifizierten kritischen Ereignissen, die auf reale Nutzungsprobleme hinweisen konnten, protokolliert und in Folge dessen auch weniger reale Nutzungsprobleme bei strikter Anwendung von IntuiBeat-F durch denselben Evaluator identifiziert wurden, soll in einem Folgeexperiment eine zusätzliche Analysesoftware zur Verbesserung der Zuordnung von Rhythmus-Peaks zu kritischen Ereignissen bereitgestellt werden.

6.2.5.3 Überprüfung der Gültigkeit

Wie erwartet, zeigte sich bei der Evaluation einer stärker intuitiv benutzbaren Software signifikant eine höhere Gültigkeit von IntuiBeat-F gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll, sowohl bei Berücksichtigung aller Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) als auch bei ausschließlicher Berücksichtigung

von durch Rhythmus-Peaks entdeckten Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F). Bei IntuiBeat-F konnte bei Berücksichtigung aller identifizierter Nutzungsprobleme (d.h. weniger strikte Anwendung) mit einer Gültigkeit von 79 % im Vergleich zum Vorgängerexperiment (siehe Teilabschnitt 6.1.4.4) 13 % weniger Gültigkeit erreicht werden (d.h. Gültigkeit von IntuiBeat-F lag im vierten Experiment bei weniger strikter Anwendung bei 94 %). Beim Nutzertest mit retrospektivem Think-Aloud-Protokoll konnte eine Gültigkeit von 64 % erreicht werden, was im Vergleich zum Vorgängerexperiment (d.h. Gültigkeit des Nutzertests lag im vierten Experiment bei weniger strikter Anwendung bei 81 %) 17 % weniger ausmacht (siehe Teilabschnitt 6.1.4.4). Bei IntuiBeat-F konnte bei Berücksichtigung ausschließlich mithilfe von Rhythmus-Peaks identifizierter Nutzungsprobleme (d.h. strikte Anwendung) mit einer Gültigkeit von 97 % im Vergleich zum Vorgängerexperiment (siehe Teilabschnitt 6.1.4.4) 1 % mehr Gültigkeit erreicht werden (d.h. Gültigkeit von IntuiBeat-F lag im vierten Experiment bei weniger strikter Anwendung bei 96 %). Obwohl unter Berücksichtigung der Studie von Koutsabasis et al. (2007) eine Gültigkeit von rund 80 % immer noch ein erstaunliches Ergebnis von IntuiBeat-F bei wenig strikter Anwendung darstellt, wird durch den Vergleich mit der Gültigkeit von nahezu 100 % bei strikter Anwendung klar (siehe Teilabschnitt 6.2.4.4), dass der Evaluator mit hoher Wahrscheinlichkeit Nutzungsprobleme unabhängig von etwaigen Rhythmus-Peaks identifizierte.

Demzufolge lässt sich bei der Berücksichtigung ausschließlich durch Rhythmus-Peaks identifizierter Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F) gegenüber der Berücksichtigung aller Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) eine Verbesserung in der Gültigkeit von circa 18 % feststellen, obwohl mithilfe der Rhythmus-Peaks nur circa 74 % aller Nutzungsprobleme und 76.60 % aller realen Nutzungsprobleme identifiziert werden konnten (siehe Teilabschnitt 6.2.4.6). Aufgrund der Tatsache, dass jedoch mithilfe der Rhythmus-Peaks nur 37 aller 50 Nutzungsprobleme identifiziert werden konnten (also bei strikter Anwendung 26 % weniger Nutzungsprobleme als bei wenig strikter Anwendung) bzw. 36 aller 47 realen Nutzungsprobleme (also bei strikter Anwendung 23.40 % weniger reale Nutzungsprobleme als bei wenig strikter Anwendung), hat sich die Gültigkeit bei strikter Anwendung von IntuiBeat-F verbessert, da diese das Verhältnis von realen zu allen gefundenen Nutzungsproblemen abbildet. Wie bereits im Vorgängerexperiment (siehe Teilabschnitt 6.1.5) wurden dementsprechend mehr nicht reale Nutzungsprobleme nicht berücksichtigt als reale Nutzungsprobleme (Verhältnis von 47 zu 50 insgesamt; verbessertes Verhältnis von 36 zu 37 durch Rhythmus-Peaks). Wie bereits unter Berücksichtigung der Ergebnisse des Vorgängerexperiments (siehe Teilabschnitt 6.1.4.4) und im Rahmen der Diskussion der Gründlichkeit im vorangegangen Teilabschnitt erwähnt (siehe Teilabschnitt 6.1.5), soll anhand einer konzeptuellen Replikation untersucht werden, inwiefern man die komplizierte Zuordnung von Rhythmus-Peaks zu entsprechenden Videostellen so vereinfachen kann, dass der Evaluator einen höheren Anteil an Rhythmus-Peaks während dem retrospektivem Interview nutzt und damit weniger kritische Ereignisse, die auf reale Nutzungsprobleme hinweisen, unabhängig von IntuiBeat-F protokolliert. Durch eine entsprechende Unterstützung des Evaluators bei der Zuordnung von Rhythmus-Peaks zu Videostellen sollte es möglich sein die übrigen 30 % der nicht genutzten Rhythmus-Peaks während des retrospektiven Interviews zu berücksichtigen und so die Gültigkeit von IntuiBeat-F systematisch zu verbessern.

Zusammenfassend können jedoch unabhängig von diesen Schlussfolgerungen, die Ergebnisse des vorliegenden fünften Experiments bezüglich der Gültigkeit bei der Evaluation einer stärker intuitiv benutzbaren Software als beeindruckend interpretiert und die in diesem Zusammenhang aufgestellten Hypothesen H2.A und H2.B entsprechend als bestätigt angesehen werden (siehe Tabelle 6.5). Aufgrund der Tatsache, dass in diesem Experiment bei der Evaluation einer stärker intuitiven und im vorangegangenen Experiment bei der Evaluation einer wenig intuitiven Software gezeigt werden konnte, dass sich die Gültigkeit bei strikter Anwendung von IntuiBeat-F verbessert, soll nun in einem Folgeexperiment untersucht werden, ob eine zusätzliche Analysesoftware zur Verbesserung der Zuordnung von Rhythmus-Peaks zu den entsprechenden kritischen Ereignissen die Gültigkeit noch weiter verbessern kann.

6.2.5.4 Überprüfung der zeitlichen Anwendungseffizienz

Wie erwartet, zeigte sich die höhere zeitliche Anwendungseffizienz von IntuiBeat-F gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll auch bei der Evaluation einer stärker intuitiv benutzbaren Software signifikant. IntuiBeat-F nahm sogar 25 % weniger Zeit für die Anwendung in Anspruch (siehe Teilabschnitt 6.2.4.5). Im Vergleich zum Vorgängerexperiment stellte dies deskriptiv sogar eine Verbesserung der zeitlichen Anwendungseffizienz um rund 5 % dar (siehe Teilabschnitt 6.1.4.5), wenn anstelle einer weniger intuitiv benutzbaren Software eine stärker intuitiv benutzbare Software verwendet wurde. Dies ist auch nicht verwunderlich, da während des fünften Experiments bei IntuiBeat-F rund 80 kritische Ereignisse bzw. 10 Nutzungsprobleme weniger und beim Nutzertest mit retrospektivem Think-Aloud-Protokoll rund 55 kritische Ereignisse bzw. 17 Nutzungsprobleme weniger als im vierten Experiment identifiziert werden mussten (siehe Teilabschnitte 6.1.4.6 und 6.2.4.6). Vergleicht man dieses Ergebnis wie beim vierten Experiment zuvor außerdem mit der Arbeit von Akers et al. (2012), liegt das Ergebnis des fünften Experiments auch in dem von ihnen ermittelten Bereich von circa 20 %, der durch eine erhöhte automatisierte Erfassung von kritischen Ereignissen wahrscheinlich ist.

Darüber hinaus kann mit hoher Wahrscheinlichkeit aufgrund der Tatsache, dass vom Evaluator lediglich rund 70 % aller Rhythmus-Peaks bei der Durchführung des retrospektiven Interviews für die Identifikation von realen Nutzungsproblemen verwendet wurden und der Evaluator stattdessen circa 57 % aller kritischen Ereignisse, die auf reale Nutzungsprobleme hinweisen konnten, unabhängig von Rhythmus-Peaks entdeckte (siehe Teilabschnitt 6.2.4.6), mit einer noch zeitlich effizienteren Durchführung gerechnet werden, wenn ein höherer Anteil an Rhythmus-Peaks beachtet wird und diese auch zu realen Nutzungsproblemen führen. Die genannten Werte liegen hierbei in einem ähnlichen Bereich wie schon im Vorgängerexperiment (siehe Teilabschnitt 6.1.5). Dementsprechend kann die in diesem Zusammenhang aufgestellte Hypothese H3 bei der Evaluation einer stärker intuitiv benutzbaren Software als bestätigt angesehen werden (siehe Tabelle 6.5). Wie bereits im Rahmen der Diskussionen zur Gründlichkeit und Gültigkeit angesprochen, soll als nächster Schritt der Evaluator in einem Folgeexperiment bei der Zuordnung von Rhythmus-Peaks zu den entsprechenden Videostellen durch eine Analysesoftware unterstützt und dabei untersucht werden, inwiefern sich die wissenschaftliche, aber auch die praktische Güte in Form von zeitlicher Anwendungseffizienz dadurch steigern kann.

6.2.6 Schlussfolgerung

Zusammenfassend kann festgehalten werden, dass die wissenschaftliche Güte von IntuiBeat-F als formative Evaluationsmethode für intuitive Benutzung hinsichtlich des Gütekriteriums *Gültigkeit* auch bei der Untersuchung einer stärker intuitiv gestalteten Software empirisch sowohl bei strikter als auch bei weniger strikter Anwendung bestätigt werden konnte (siehe Tabelle 6.5). Darüber hinaus konnte IntuiBeat-F auch praktische Güte aufgrund seiner im Vergleich zum Nutzertest mit retrospektivem Think-Aloud-Protokoll höheren zeitlichen Anwendungseffizienz empirisch zugesprochen werden (siehe Tabelle 6.5). Konfundierungen durch Unterschiede in der Vorerfahrung bei der Nutzung von CUIs können mit hoher Wahrscheinlichkeit ausgeschlossen werden. Die Effekt- und Teststärken lagen dabei überwiegend im oberen Bereich. Wie bereits im Vorgängerexperiment konnte jedoch die wissenschaftliche Güte von IntuiBeat-F bezüglich des Gütekriteriums der Gründlichkeit nicht bestätigt werden, wenn die Anwendung von IntuiBeat-F strikt erfolgte und damit nur Nutzungsprobleme als Grundlage für die Bestimmung der Gründlichkeit genutzt wurden, die mithilfe von Rhythmus-Peaks gefunden wurden (siehe Tabelle 6.5).

Tabelle 6.5. *Übersicht der mithilfe des fünften Experiments bestätigten Hypothesen im Zuge der Meta-Evaluation von IntuiBeat-F (KEs: kritische Ereignisse; RP-KEs: mit Rhythmus-Peaks assoziierte kritische Ereignisse).*

Hypothese	Experiment 5
(H1) Überprüfung der Gründlichkeit:	
- (A) Alle KEs	✓
- (B) Nur RP-KEs	✗
(H2) Überprüfung der Gültigkeit:	
- (A) Alle KEs	✓
- (B) Nur RP-KEs	✓
(H3) Überprüfung der zeitlichen Anwendungseffizienz	✓

Da im Vergleich zum Vorgängerexperiment, in dem eine weniger intuitiv benutzbare Software untersucht wurde, ein ähnliches Ausmaß an nicht berücksichtigten Rhythmus-Peaks erkennbar war, und vom selben Evaluator kritische Ereignisse überwiegend unabhängig von Rhythmus-Peaks abgeleitet wurden, was zu Lasten der Gültigkeit (d.h. Protokollierung von kritischen Ereignissen unabhängig von Rhythmus-Peaks hinter denen sich lediglich Bewegungsfehler und damit keine realen Nutzungsprobleme verbargen), der Gründlichkeit (d.h. reale Nutzungsprobleme wurden entweder übersehen, weil sie nicht durch Rhythmus-Peaks erkannt wurden oder der Evaluator sie vorsätzlich nicht zugeordnet hat) und der zeitlichen Anwendungseffizienz (d.h. Rhythmus-Peaks wurden nicht vollständig genutzt und stattdessen mit wahrscheinlich höherem zeitlichen Aufwand kritische Ereignisse ohne Unterstützung identifiziert) bei der Evaluation einer stärker intuitiv benutzbaren Software ging, soll in zwei Folgeexperimenten (d.h. Untersuchungsgegenstand ist einmal eine weniger intuitiv benutzbare und einmal eine stärker intuitiv benutzbare Software) dem Evaluator eine zusätzliche Analysesoftware für das retrospektive Interview bereitgestellt werden. Mithilfe dieser Software soll die manuelle Zuordnung der Rhythmus-Peaks zu den entsprechenden Videostellen bzw. kritischen Ereignissen entfallen und durch die erhöhte

Automatisierung verifiziert werden, ob damit Gründlichkeit (d.h. Entdeckung eines hohen Anteils an realen Nutzungsproblemen), Gültigkeit (d.h. Vermeidung der Identifikation nicht realer Nutzungsprobleme) und zeitliche Anwendungseffizienz (d.h. Vermeidung der zeitlich aufwendigen Protokollierung Rhythmus-Peaks-unabhängiger kritischer Ereignisse) gesteigert werden können, wenn ausschließlich Rhythmus-Peaks für die Identifikation kritischer Ereignisse (d. h. strikte Anwendung von IntuiBeat-F) herangezogen werden.

6.3 Experiment 6

Das in diesem Abschnitt vorgestellte sechste Experiment untersuchte, wie die beiden Experimente zuvor, inwiefern IntuiBeat-F wissenschaftliche Güte bezüglich der formalen Hauptgütekriterien *Gründlichkeit* und *Gültigkeit*, sowie der *zeitlichen Anwendungseffizienz*, als für das Projekt 3D-GUIde wichtiger Aspekt praktischer Güte, attestiert werden kann. Da in den letzten beiden Experimenten das Gütekriterium der Gründlichkeit bei strikter Anwendung von IntuiBeat-F nicht bestätigt werden konnte (siehe Tabellen 6.3 und 6.5) und mit hoher Wahrscheinlichkeit angenommen werden kann (siehe Diskussionen in den Teilabschnitten 6.1.5 und 6.2.5), dass sich die manuelle Zuordnung der Rhythmus-Peaks zu den Videostellen durch den Evaluator negativ auf die wissenschaftliche Güte und die zeitliche Anwendungseffizienz auswirkte (z.B. es können aufgrund der Tabellendarstellung leicht Rhythmus-Peaks vergessen werden), wurde im Zuge des sechsten Experiments eine zusätzliche Analysesoftware im Rahmen des retrospektiven Interviews eingesetzt.

Unter Berücksichtigung der bisherigen IntuiBeat-Software und der neuen Analysesoftware kam es im Rahmen dieses Experiments zu einem Vergleich mit einem als Quasi-Außenkriterium fungierenden Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der Nutzung eines CUIs, nämlich des CAD-Programms *Tinkercad* (Autodesk, 2017b). Es wurde sich für ein CUI entschieden, da die wissenschaftliche Güte und zeitliche Anwendungseffizienz von IntuiBeat-F aufgrund der Anforderungen des Projekts 3D-GUIde zunächst im Bereich von CUIs unter Berücksichtigung dieser Anpassung erneut bestätigt werden mussten. Beim dem als Untersuchungsgegenstand verwendeten CUI handelte es sich, auf Basis einer Experteneinschätzung, um eine eher weniger intuitiv benutzbare Software. Das sechste Experiment betrachtete dementsprechend wie die beiden Experimente zuvor auch die zweite Forschungsfrage und den formativen Aspekt der dritten Forschungsfrage dieser Arbeit.

6.3.1 Überprüfung der Gründlichkeit und Gültigkeit

Aufgrund der Tatsache, dass es es beim sechsten Experiment lediglich um eine konzeptuelle Replikation des vierten Experiments handelte, wobei ein anderes CUI und eine zusätzliche Analysesoftware für das retrospektive Interview eingesetzt wurden, wurden auch die gleichen Hypothesen wie bei den Vorgängerexperimenten aufgestellt, weswegen an dieser Stelle auf den entsprechenden Teilabschnitt 6.1.1 des vierten Experiments verwiesen werden soll.

6.3.2 Überprüfung der zeitlichen Anwendungseffizienz

Auch bezüglich der zeitlichen Anwendungseffizienz wurde die entsprechende Hypothese aus den beiden Vorgängerexperimenten übernommen, weswegen an dieser Stelle auf den entsprechenden Teilabschnitt 6.1.2 des vierten Experiments verwiesen werden soll.

6.3.3 Methode

6.3.3.1 Teilnehmer

Für das fünfte Experiment wurden 29 Versuchspersonen über das Probandensystem des Instituts für Mensch-Computer-Medien an der Universität Würzburg rekrutiert. Aufgrund der Tatsache, dass bei einer Person die Rhythmusaufzeichnung nicht vollständig war (d.h. Rhythmusaufzeichnung bei einer Teilaufgabe vergessen) und bei vier Personen die Videoaufzeichnung vergessen wurde, mussten diese Datensätze von der Datenauswertung ausgeschlossen werden. Demzufolge konnten für die Meta-Evaluation von IntuiBeat-F 24 Versuchspersonen berücksichtigt werden, welche alle rechtsfüßig (d.h. der rechte Fuß stellte den dominanten Fuß dar) waren. Die Versuchspersonen setzten sich dabei aus 17 Frauen und sieben Männern zusammen. Das Durchschnittsalter betrug 21.21 Jahre (SD = 3.16). Es handelte sich bei allen Teilnehmern um Studierende der Julius-Maximilians-Universität Würzburg, wobei sechs Personen Mensch-Computer-Systeme (25 %) und 18 Personen Medienkommunikation (75 %) studierten. Alle Versuchsteilnehmer wurden über das Probanden-System des Instituts Mensch-Computer-Medien über eine gesonderte Mail darauf hingewiesen, für den Versuch flache Sportschuhe zu tragen, um eine möglichst problemlose Rhythmuseingabe über das USB-Fußpedal zu ermöglichen. Für die Teilnahme an der Untersuchung bekam jede Versuchsperson eine halbe Versuchspersonenstunde gutgeschrieben. Die mit einem TFQ gemessene Vorerfahrung der Versuchspersonen bezüglich der Nutzung von CUIs betrug im Durchschnitt .06 (SD = .17) bei einem Maximum von 6 und lag damit erwartungsgemäß im unteren Bereich. Alle Versuchspersonen besaßen damit eine geringe Vorerfahrung mit CUIs. Alle Versuchspersonen gaben an, am Experiment freiwillig teilzunehmen.

6.3.3.2 Versuchsdesign

Für die Beantwortung der zweiten Forschungsfrage und des formativen Aspekts der dritten Forschungsfrage wurde das gleiche Experimentaldesign wie zuvor beim vierten Experiment genutzt. Für genauere Informationen bezüglich der unabhängigen und abhängigen Variablen wird daher auf den entsprechenden Absatz 6.1.3.1 des vierten Experiments verwiesen.

6.3.3.3 Versuchsmaterialien und Maße

Die im sechsten Experiment verwendeten Versuchsmaterialien und Maße unterschieden sich nur einigen Punkten von den im vierten und fünften Experiment verwendeten Versuchsmaterialien und Maßen, weswegen in diesem Teilabschnitt auch nur auf diese einge-

gangen wird und für die übernommenen Versuchsmaterialien und Maße auf den entsprechenden Absatz 6.1.3.3 des vierten Experiments verwiesen wird.

Untersuchungsgegenstand der formativen Evaluation: Tinkercad

Als Untersuchungsgegenstand für die Meta-Evaluation von IntuiBeat-F kam die Online-CAD-Software *Tinkercad* von Autodesk (2017b) als weniger intuitiv benutzbare Software auf Basis einer qualitativen Experteneinschätzung ($N_{Experte}$= 5; Vorgehen: siehe entsprechenden Absatz innerhalb des Teilabschnitts 5.1.4.3) im Rahmen des sechsten Experiments zum Einsatz (siehe Abbildung 6.10). Es wurde für die Meta-Evaluation, wie beim vierten Experiment, auf ein CUI als Untersuchungsgegenstand gesetzt, da mit der Einführung der neuen Analysesoftware zur Unterstützung des retrospektiven Interviews erneut sichergestellt werden musste, inwiefern die Methode auch unter Berücksichtigung dieser Anpassung in dem für das Projekt 3D-GUIde relevanten Bereich der CUIs wissenschaftlich tragfähige Ergebnisse auf eine zeitlich effiziente Art und Weise erzielen kann.

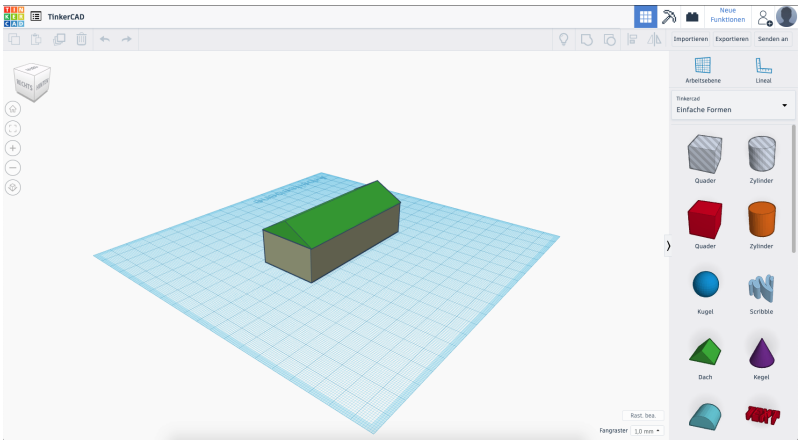


Abbildung 6.10. Die im Rahmen des sechsten Experiments als Untersuchungsgegenstand verwendete Online-CAD-Software *Tinkercad* von Autodesk (2017b).

Es wurde sich als Untersuchungsgegenstand für ein klassisches CAD-Programm entschieden, da unter Berücksichtigung der für das Experiment zur Verfügung stehenden Studierendenstichprobe zu erwarten ist, dass Tinkercad von den Studierenden überwiegend als weniger intuitiv benutzbare Softwareanwendung eingestuft wird (d.h. weniger intuitiv benutzbare Softwareanwendungen sollten logischerweise auch zu mehr Nutzungsproblemen führen) und so beide formative Evaluationsmethoden ihr Potential zur Entdeckung von Nutzungsproblemen ausspielen können. Darüber hinaus sollte im Rahmen des Projekts 3D-GUIde auch eine Online-CAD-Software evaluiert werden, weswegen sich Tinkercad auch aus diesem Grund als Untersuchungsgegenstand anbot.

Aufgrund der Tatsache, dass es sich bei Tinkercad um ein klassisches CAD-Programm handelt, sollte dieses immer die von LaViola et al. (2017) beschriebenen domänenübergreifenden, grundlegenden Nutzeraufgaben Selektion, Manipulation, Navigation, Zeicheneingaben und Systemsteuerung unterstützen, weswegen vier experimentelle Aufgaben (siehe Anhang B.6.1) für das sechste Experiment entsprechend aus diesen Bereichen gewählt wurden. Auf diese Weise konnten wie bei den anderen CUI-Experimenten intuitiv benutzbare

3D-CUI-Interaktionslösungen künftig als Interaktionspatterns auf Basis der Ergebnisse abgeleitet werden. Sie wurden von den Versuchspersonen in einer festen Reihenfolge bearbeitet. Alle Aufgaben wurden dabei in ein fiktives 3D-Druck-Szenario eingebettet. Hierbei mussten die Versuchspersonen als Erstes ein bereits erstelltes 3D-Objekt (d.h. Hausdach) auf ihre Arbeitsfläche einfügen, auf der sich bereits ein Quader befand (erste Aufgabe). Anschließend mussten die Versuchspersonen das Dach so drehen, dass es später der Länge nach auf dem Quader platziert werden konnte (zweite Aufgabe). Daraufhin mussten die Versuchspersonen die Größe des Daches so anpassen, dass das Dach den Quader später vollständig bedecken konnte (dritte Aufgabe). Abschließend mussten die Versuchspersonen das Dach auf dem Quader positionieren und dieses gegebenenfalls in der Größe und Orientierung anpassen (vierte Aufgabe).

Formative Evaluation intuitiver Benutzung mit IntuiBeat-F

Um die in der Marker-Datei enthaltenen Rhythmus-Peaks nicht manuell den Videostellen bzw. den kritischen Ereignissen während des retrospektiven Interviews zuordnen zu müssen, wurde, in Ergänzung zur in Absatz 6.1.3.3 des vierten Experiments beschriebenen Vorgehensweise bei der formativen Evaluation intuitiver Benutzung mithilfe von IntuiBeat-F, eine zusätzliche Analysesoftware für die automatisierte Zuteilung eingesetzt. Diese wurde dem Evaluator ebenfalls als JAR-Datei auf seinem Evaluator-PC bereitgestellt (siehe Absatz „Apparatur“ innerhalb des Teilabschnitts 6.3.3.3).

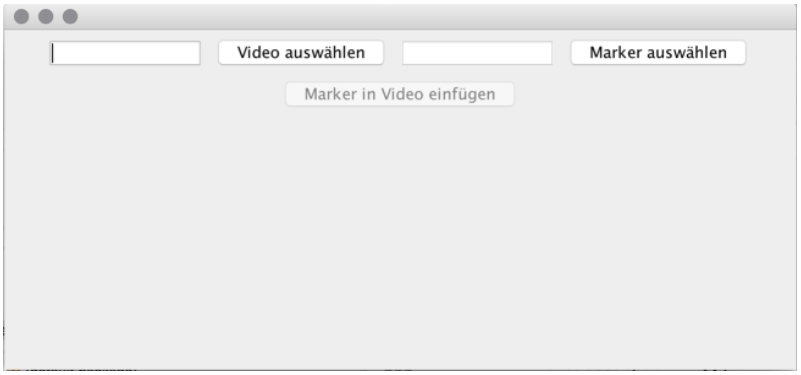


Abbildung 6.11. Die im Rahmen des vierten und fünften Experiments eingesetzte Analysesoftware, welche Rhythmus-Peaks automatisch in der Bildschirmaufzeichnung markiert.

Der Evaluator musste nach dem Start der Software zunächst die Bildschirmaufzeichnung einer Aufgabe von einer Versuchsperson (z.B. „vp01_1.mp4“) mit einem Klick auf den Button „Video auswählen“ auswählen, die er zuvor während dem Experiment aufgenommen hatte (siehe Abbildung 6.11). An dieser Stelle sei darauf hingewiesen, dass die Analysesoftware lediglich MP4 als Eingabeformat akzeptiert. Für alle weiteren Formate muss der Code entweder angepasst oder die Videoaufzeichnung entsprechend im Vorfeld umkodiert werden. Im Anschluss wählte der Evaluator die entsprechende Marker-Datei durch Klick auf den Button „Marker auswählen“ aus, deren Rhythmus-Peaks er gerne automatisch den entsprechenden Zeitpunkten in der Bildschirmaufzeichnung zugeordnet hätte. Abschließend betätigte er die Zuordnung durch Klick auf den Button „Marker in Video einfügen“ (siehe Abbildung 6.11).

Die Implementierung der Analysesoftware erfolgte wie bereits die der IntuiBeat-Software in Java, wobei zusätzlich die Websprachen HTML, CSS und JavaScript in Form von jQuery (Resig, 2017) verwendet wurden. Um die Rhythmus-Peaks dem Fortschrittsbalken der Bildschirmaufzeichnung in Form von Markierungen hinzufügen zu können, wurde das quelloffene Projekt VideoJS-Markers (Spchuang, 2016) verwendet, welches auf dem ebenfalls quelloffenen Projekt Video.js (Heffernan, 2013) aufbaut. Letzteres erlaubt es Videos per HTML 5 ablaufen zu lassen. Um mithilfe der Analysesoftware nun die Rhythmus-Peaks der Bildschirmaufzeichnung zuzuordnen, wurde dem Evaluator neben der eigentlichen Analysesoftware zusätzlich ein HTML-Template (d.h. „template.html“) bereitgestellt, welches über eine Variable für den Pfad des Videos und eine Variable für eine Liste von Rhythmus-Peaks verfügte. Durch Klicken auf den Button „Marker in Video einfügen“ wurde diesen Variablen dann dynamisch während der Laufzeit der Analysesoftware der Pfad des ausgewählten Videos und die Inhalte aus der angegeben Marker-Datei zugewiesen, um als Resultat eine wie das Video benannte HTML-Seite (z.B. „vp01_1.html“) auf Basis des Templates erzeugen zu können.

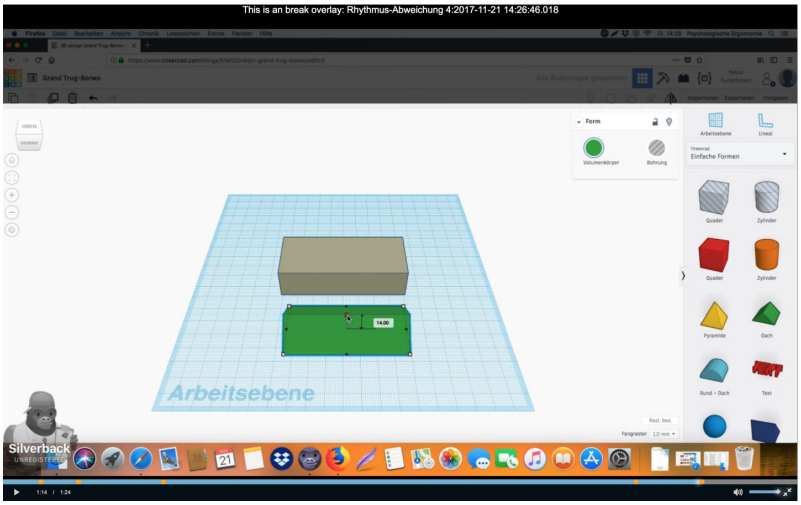


Abbildung 6.12. Beispiel für eine mit der Analysesoftware generierte HTML-Datei, mit der die Rhythmus-Peaks automatisch in der Bildschirmaufzeichnung markiert wurden.

Diese HTML-Seite enthielt das angegebene Video, dem die Rhythmus-Peaks entsprechend zeitlich zugeordnet wurden (siehe Abbildung 6.12). Um die Analysesoftware dem Evaluator als JAR-Datei ohne größeren Einrichtungsaufwand bereitzustellen, musste das HTML-Template im selben Ordner wie die JAR-Datei selbst platziert werden, da man ansonsten noch einen weiteren Dialog zur Bestimmung des Pfads des HTML-Templates hätte einbauen müssen. Es wurde sich für ein auf HTML 5 basierendes Verfahren entschieden, da man so unabhängig von Betriebssystem und installierter Software sicherstellen konnte, dass die Bildschirmaufzeichnung mit den gesetzten Rhythmus-Peaks immer abspielbar und als HTML theoretisch von überall zugreifbar ist. Der Evaluator konnte dann mit der generierten HTML-Datei das retrospektive Interview für die entsprechende Aufgabe durchführen, so wie es bereits in Absatz „Formative Evaluation intuitiver Benutzung mit IntuiBeat-F“ des Teilabschnitts 6.1.3.3 des vierten Experiments beschrieben ist, nur dass die automatische Zuordnung der Rhythmus-Peaks zu den dazugehörigen Videostellen durch die neue Analysesoftware etwaiges Fehlerpotential minimierte und damit eine

höhere Gründlichkeit potentiell ermöglichte. Die Vorgehensweise und die Testumgebung für die formative Evaluation intuitiver Benutzung mithilfe der IntuiBeat-Software wurde jedoch nicht verändert und ist somit mit der entsprechenden Beschreibung innerhalb des vierten Experiments identisch (siehe Absatz „Formative Evaluation intuitiver Benutzung mit IntuiBeat-F“ des Teilabschnitts 6.1.3.3).

Vorerfahrung bei der Nutzung von CUIs

Die Vorerfahrung der Versuchspersonen bezüglich der CAD-Software *Tinkercad* wurde mit einem hierfür konstruierten TFQ als Kontrollvariable erhoben (KV_Vorerfahrung), welcher wie in den Experimenten zuvor, in Papierform administriert wurde. Als Items für diesen Fragebogen wurden lediglich „Tinkercad“ und „Andere CAD-Software?“ genutzt, wobei letztere eine Wildcard darstellte und die Angabe einer beliebigen bekannten CAD-Software erlaubte. Es wurde bei der Konstruktion des TFQ im Vergleich zu einigen TFQs aus den vorigen Experimenten auf die Nennung konkreter CUIs als Beispiele verzichtet, da bei vorigen Studien, bei denen ebenfalls die in diesem Experiment als Stichprobe fungierenden Studierenden der Medienkommunikation und Mensch-Computer-Systeme eingesetzt wurden, üblicherweise eh nur ein bestimmtes CUI bekannt war und genutzt wurde. Dementsprechend war es ausreichend, die Expertise der Studierenden mit der Wildcard entsprechend zu erfassen.

Wie bei den anderen konstruierten TFQs, beurteilten die Versuchspersonen diese beiden Items dann bezüglich deren Nutzungshäufigkeit (Wertebereich: 0 - 6) und des genutzten Funktionsumfangs (Wertebereich: 0 - 4). Im Anschluss wurde dann für jede Dimension (d.h. Nutzungshäufigkeit und Funktionsumfang) ein Mittelwert gebildet und daraufhin ein Gesamtmittelwert über beide Mittelwerte als Operationalisierung der *Vorerfahrung bei der Nutzung von CUIs* berechnet (Wertebereich: 0 - 5). Die Entscheidung für eine Mittelwertbildung anstelle einer Summenberechnung wurde im Rahmen des ersten Experiments ausführlich begründet, weswegen an dieser Stelle lediglich darauf verwiesen wird (siehe Absatz „Vorerfahrung bei der Nutzung von CUIs“ des Teilabschnitts 5.1.4.3). Da kein einheitlicher TFQ existiert und dieser jeweils für die getesteten Softwareanwendungen erstellt werden musste, ist der in diesem Experiment genutzte TFQ vollständig in Anhang B.6.2 dieser Arbeit zu finden.

Apparatur

Die im Rahmen des sechsten Experiments verwendete Apparatur unterschied sich nur in einigen Punkten von der im zweiten Experiment genutzten Apparatur, die lediglich einen Versuchspersonen-PC und Analyse-PC nutzte. Ein Unterschied bestand darin, dass als Analyse-PC ein 15 Zoll Apple MacBook Pro (MacOS High Sierra, 2.3 GHz Quad-Core Intel i7, 4 GB Arbeitsspeicher, 1536 MB Intel HD Graphics 4000, Auflösung 2880 x 1800 Pixel) und als Versuchspersonen-PC ein 21.5 Zoll Apple iMac (MacOS High Sierra, 2.7 GHz Quad-Core Intel i7, 8 GB Arbeitsspeicher, 512 MB AMD Radeon HD 6770M, Auflösung 1920 x 1080 Pixel) zum Einsatz kamen. Als Eingabegeräte für den iMac wurde eine MagicMouse 1 und ein Apple Magic Keyboard verwendet. Der Evaluator nutzte für die Benutzung des Analyse-PCs lediglich das eingebaute Trackpad. Darüber hinaus kam anstelle

von Captura die Bildschirmaufzeichnungssoftware Silverback 3 von Clearleft Ltd. (2016) zum Einsatz. Als Browser für die Nutzung von Tinkercad von Autodesk (2017b) wurde Mozilla Firefox in der Version 52.4.1 genutzt. Die restliche Apparatur war identisch zum zweiten Experiment, weswegen für weitere Details auf den entsprechenden Teilabschnitt des zweiten Experiments (siehe Absatz „Apparatur“ des Teilabschnitts 5.2.3.3) verwiesen werden soll.

6.3.3.4 Versuchsdurchführung

Die Versuchsdurchführung des sechsten Experiments war im Vergleich zum vierten Experiment mit der Ausnahme identisch, dass von den Versuchspersonen anstelle der Regalplanungssoftware *IPO.Rack* mit der CAD-Software *Tinkercad* gearbeitet wurde, weswegen an dieser Stelle für die Beschreibung der Versuchsdurchführung auf den entsprechenden Teilabschnitt 6.1.3.4 des vierten Experiments verwiesen werden soll. Es sei hier lediglich darauf hingewiesen, dass der Evaluator im sechsten Experiment nicht wie in den beiden Vorgängerexperimenten nur die entsprechende Software bzw. Website öffnen musste, sondern auch zusätzlich das für die Aufgaben erforderliche 3D-Modell in Tinkercad selbst geladen werden musste (Datei „TinkercadExperiment6.stl“).

6.3.3.5 Statistische Auswertung

Die statistische Auswertung des sechsten Experiments war identisch mit der statistischen Auswertung der beiden Vorgängerexperimente, weswegen für genauere Informationen auf den entsprechenden Teilabschnitt 6.1.3.5 des vierten Experiments verwiesen werden soll. Da die *Vorerfahrung bei der Nutzung von CUIs*, die als Kontrollvariable erhoben wurde, einen ungewollten Einfluss auf die Gründlichkeit und Gültigkeit der verglichenen Methoden haben könnte, musste eine derartige Konfundierung vor der Überprüfung der Gründlichkeit und Gültigkeit von IntuiBeat-F ausgeschlossen werden. Hierzu wurden t-Tests für unabhängige Stichproben bezüglich der Vorerfahrung, sowie Pearson-Produkt-Moment-Korrelationen zwischen der Vorerfahrung und den abhängigen Variablen innerhalb der beiden Ausprägungen der unabhängigen Variablen gerechnet (siehe Teilabschnitt 6.3.4.2). Da in beiden Fällen keine signifikanten Ergebnisse festgestellt werden konnten (d.h. kein ungewollter Einfluss der Vorerfahrung auf Gründlichkeit und Gültigkeit), wurde von einer univariaten Kovarianzanalyse (ANCOVA) abgesehen (Döring & Bortz, 2016; Field, 2017) und sich stattdessen für eine Datenauswertung mithilfe von t-Tests für unabhängige Stichproben ohne Berücksichtigung der Vorerfahrung bei der Nutzung von CUIs als Kovariate entschieden (siehe Abschnitt 6.3.4). Die Überprüfung der statistischen Voraussetzungen der Pearson-Produkt-Moment-Korrelationen und der t-Tests werden im folgenden Ergebnisteil berichtet (siehe Teilabschnitt 6.3.4.1).

6.3.4 Ergebnisse

Im folgenden Abschnitt werden die Ergebnisse bezüglich der in den Teilabschnitt 6.3.1 und 6.3.2 beschriebenen Hypothesen deskriptiv und inferenzstatistisch berichtet. Vor der

eigentlichen Datenanalyse wird zunächst auf die Überprüfung der statistischen Voraussetzungen eingegangen. Dabei wurde bei allen abhängigen Variablen und Kontrollvariablen ein metrisches Skalenniveau angenommen.

6.3.4.1 Überprüfung der statistischen Voraussetzungen

Überprüfung von Ausreißern

Das Vorgehen bei der Ausreißeranalyse des sechsten Experiments war identisch mit dem Vorgehen im vierten Experiment, weswegen für genauere Informationen auf den entsprechenden Teilabschnitt 6.1.4.1 des vierten Experiments verwiesen werden soll. Es mussten auf diese Weise bei keiner der analysierten abhängigen Variablen und Kontrollvariablen Werte ausgeschlossen werden.

An dieser Stelle ist jedoch zusätzlich anzumerken, dass aufgrund der geringen Varianz von KV_Vorerfahrung ($M = .06$, $SD = .17$) und der damit verbundenen Tatsache, dass nahezu alle Versuchspersonen keine Vorerfahrung bei der Nutzung von CUIs aufwiesen und die angegebenen Werte damit nahezu das Minimum des TFQ (d.h. Minimum des TFQ lag bei 0) repräsentierten, in diesem Fall kein modifizierter z-Wert berechnet werden konnte (d.h. Dividende durch Null). Stattdessen wurde anhand eines Boxplots von KV_Vorerfahrung überprüft, inwiefern darin univariate Ausreißer zu finden sind (siehe Howell, 2009). Mit diesem Verfahren konnten drei Ausreißer ($\text{Ausreißer}_1 = .50$, $\text{Ausreißer}_2 = .50$, $\text{Ausreißer}_3 = .50$) festgestellt werden, welche aber aufgrund der geringen Stichprobengröße und der Tatsache, dass ansonsten die Vorerfahrung konstant bei Null gewesen wäre, nicht ausgeschlossen wurden.

Überprüfung der Voraussetzung der Normalverteilung

Die univariate Normalverteilung der abhängigen Variablen und Kontrollvariablen wurde für jede Ausprägung der unabhängigen Variable mittels Kolmogorov-Smirnov-Tests ($p \geq .05$, siehe Field, 2017) und Sichtprüfung anhand eines Q-Q-Diagramms geprüft. Dabei konnten auf Basis dieser beiden Kriterien bei AV_Gründlichkeit_WenigerStrikt bzw. AV_Gründlichkeit_Strikt im Rahmen der Überprüfung der Gründlichkeit (Hypothese H1), bei der AV_Gültigkeit_WenigerStrikt und der AV_Gültigkeit_Strikt im Rahmen der Überprüfung der Gültigkeit (Hypothese H2) und bei der KV_Vorerfahrung keine Normalverteilung in beiden Gruppen festgestellt werden. Aufgrund der Tatsache, dass ein ungepaarter t-Test bei etwa gleich großen Gruppen robust gegenüber Verletzungen der Normalverteilungsannahme ist (Glass et al., 1972; Pagano, 2006; Salkind, 2010; Wilcox, 2011), wurde sich für eine eindeutige Interpretation und Vergleichbarkeit der Ergebnisse (d.h. Mittelwerte statt Medianen) der verschiedenen Experimente gegen Transformationen und entsprechende nonparametrische Verfahren zur Überprüfung der Hypothesen (d.h. H1 und H2) entschieden.

Überprüfung der Voraussetzung der Homoskedastizität

Zur Überprüfung der Homoskedastizität zwischen den Ausprägungen der unabhängigen Variable kamen Levene-Tests zum Einsatz (siehe Field, 2017), welche im Rahmen der Überprüfung der Gründlichkeit (Hypothese H1), der Überprüfung der Gültigkeit (Hypothese H2) und der Überprüfung der zeitlichen Anwendungseffizienz (Hypothese H3) gerechnet und bei allen abhängigen Variablen und Kontrollvariablen ($p \geq .05$, siehe Field, 2017) außer bei der AV_Gültigkeit_WenigerStrikt und der AV_ZeitlicheAnwendungseffizienz Varianzhomogenität bestätigen konnten. Die Freiheitsgrade für die dazugehörigen t-Tests im Rahmen der Überprüfung der Gültigkeit (Hypothese H2) und der Überprüfung der zeitlichen Anwendungseffizienz (Hypothese H3) wurden dementsprechend korrigiert.

6.3.4.2 Überprüfung einer möglichen Konfundierung durch Vorerfahrung bei der Nutzung von CUIs

Die Vorerfahrung bei der Nutzung von CUIs (Wertebereich: 0 - 5) unterschied sich nicht signifikant zwischen den beiden Ausprägungen der unabhängigen Variable (Nutzertest mit retrospektivem Think-Aloud-Protokoll: $M = .04$; $SD = .14$; IntuiBeat-F: $M = .08$, $SD = .20$), $t(22) = -.60$, $p = .557$, d = .23. Laut J. Cohen (1988) kann dieser Effekt als mittel interpretiert werden ($d \leq .5$). Eine konservative post-hoc Analyse der Teststärke bezüglich KV_Vorerfahrung mithilfe von G*Power (Faul et al., 2009) mit $df = 22$ und einer angenommenen geringen Effektstärke (d = .23) ergab lediglich eine geringe Teststärke (1 - $\beta = .10$) im Zuge der Auswertung der Vorerfahrung bei der Nutzung von CUIs. Um jedoch eine ausreichend große Power (1 - $\beta \geq .80$) beim festgestellten Effekt erzielen zu können, wären pro Gruppe 486 Versuchspersonen nötig gewesen, was aufgrund des straffen Zeitplans im Anwenderprojekt, des dortigen Fokus auf qualitative Ergebnisse, des Verständnisses der Meta-Evaluation von IntuiBeat-F als Nebenprodukt und der personellen Einschränkungen im Projekt 3D-GUIde nicht möglich gewesen wäre. Auch bei Annahme einer großen Effektstärke (d = .8) hätten immerhin noch 26 Versuchspersonen pro Gruppe getestet werden müssen, was im Hinblick auf die Einschränkungen durch das Anwenderprojekt nicht notwendig erschien.

Da die erhobene Vorerfahrung bei der Nutzung von CUIs dennoch einen ungewollten Einfluss auf den Vergleich der beiden Ausprägungen der unabhängigen Variable haben könnte, wurde für die Kontrollvariable mithilfe von Pearson-Produkt-Moment-Korrelationen sichergestellt, dass kein Zusammenhang zwischen dieser und den abhängigen Variablen innerhalb der beiden Ausprägungen der unabhängigen Variable besteht ($p > .05$). Die Kontrollvariable KV_Vorerfahrung wurde dementsprechend nicht als Kovariate in der statistischen Auswertung berücksichtigt, da kein linearer Zusammenhang zwischen dieser und den abhängigen Variablen bestand (siehe Döring & Bortz, 2016; Field, 2017).

6.3.4.3 Überprüfung der Gründlichkeit

Wie erwartet, lag die Gründlichkeit (%) von IntuiBeat-F ($M = .27$, $SD = .09$) signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll ($M = .18$, $SD = .09$), wenn alle mit IntuiBeat-F abgeleiteten Nutzungsprobleme (d.h. weniger strikte

Anwendung von IntuiBeat-F) berücksichtigt wurden (H1.A), $t(22) = -2.56$, $p = .018$, d = 1. Es war laut J. Cohen (1988) ein großer Effekt feststellbar (d $\geq$ .8). Aufgrund der Tatsache, dass alle mit IntuiBeat-F abgeleiteten Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) auch mithilfe des Algorithmus abgeleitet werden konnten (d.h. strikte Anwendung von IntuiBeat-F), konnte dementsprechend auch bei der strikten Anwendung ein statistisch identisches Ergebnis wie bei der weniger strikten Anwendung erzielt werden (H1.B).

6.3.4.4 Überprüfung der Gültigkeit

Wie erwartet, lag die Gültigkeit (%) von IntuiBeat-F ($M = .88$, $SD = .15$) signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll ($M = .51$, $SD = .11$), wenn alle mit IntuiBeat-F abgeleiteten Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) berücksichtigt wurden (H2.A), $t(20.54) = -6.94$, $p < .001$, d = 2.81. Es war laut J. Cohen (1988) ein großer Effekt feststellbar (d $\geq$ .8). Auch wenn ausschließlich die auf Basis des Algorithmus von IntuiBeat-F abgeleiteten Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F) berücksichtigt wurden (H2.B), lag die Gültigkeit, wie erwartet, von IntuiBeat-F ($M = .90$, $SD = .15$) signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll ($M = .51$, $SD = .11$), $t(22) = -7.41$, $p < .001$, d = 2.97. Es war laut J. Cohen (1988) hier ebenfalls ein großer Effekt feststellbar (d $\geq$ .8).

6.3.4.5 Überprüfung der zeitlichen Anwendungseffizienz

Wie erwartet, lag die zeitliche Anwendungseffizienz (s) von IntuiBeat-F in Form einer geringeren durchschnittlichen Problemidentifikationszeit für reale Nutzungsprobleme ($M = 421.96$, $SD = 139.36$) signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll ($M = 721.33$, $SD = 273.00$), $t(16.37) = 3.38$, $p = .003$, d = 1.38. Es war laut J. Cohen (1988) ein großer Effekt feststellbar (d $\geq$ .8).

6.3.4.6 Explorative Datenanalyse

Wie bereits in Teilabschnitt 6.1.3.3 des vierten Experiments angesprochen, wurden zusätzlich eine Reihe von Metriken für eine explorative Datenanalyse berechnet, mit deren Hilfe verstanden werden kann, warum es zu keinem Unterschied bezüglich Gründlichkeit (siehe Teilabschnitt 6.3.4.3) und zu einem Unterschied bezüglich Gültigkeit (siehe Teilabschnitt 6.3.4.4) bei der strikten und weniger strikten Anwendung von IntuiBeat-F gekommen ist. Im Folgenden sollen diese Metriken nun berichtet werden.

Tabelle 6.6. *Metriken zur explorativen Datenanalyse von IntuiBeat-F und des Nutzertests mit retrospektivem Think-Aloud-Protokoll im Rahmen des sechsten Experiments (NPs: Nutzungsprobleme; KEs: kritische Ereignisse; RP-KEs: mit Rhythmus-Peaks assoziierte kritische Ereignisse; Alle NPs aus KEs ≈ weniger strikte Anwendung von IntuiBeat-F; Nur NPs aus RP-KEs ≈ strikte Anwendung von IntuiBeat-F).*

Metrik	Experiment 6	
	IntuiBeat-F	Nutzertest
Trefferrate NPs (%):		
- Alle NPs aus KEs	93.94	88.24
- Nur NPs aus RP-KEs	100	-
Trefferrate reale NPs (%):		
- Alle NPs aus KEs	84.85	26.47
- Nur NPs aus RP-KEs	93.33	-
Fehlalarmrate NPs (%):		
- Alle NPs aus KEs	6.06	11.74
- Nur NPs aus RP-KEs	0	-
Fehlalarmrate reale NPs (%):		
- Alle NPs aus KEs	15.15	73.53
- Nur NPs aus RP-KEs	6.67	-
Anzahl NPs:		
- Alle NPs aus KEs	10	6
- Nur NPs aus RP-KEs	10	-
Anzahl realer NPs:		
- Alle NPs aus KEs	9	1
- Nur NPs aus RP-KEs	9	-
Anteil realer NPs an NPs (%):		
- Alle NPs aus KEs	90	16.67
- Nur NPs aus RP-KEs	90	-

Explorative Datenanalyse von IntuiBeat-F (weniger strikte Anwendung)

Mithilfe von IntuiBeat-F konnten über alle Versuchspersonen hinweg 33 kritische Ereignisse erkannt werden, wovon unter Berücksichtigung der in der Abbildung 6.2 in Teilabschnitt 6.1.3 beschriebenen Arbeitsschritte zur Ableitung von Nutzungsproblemen (Burmester, 2016) und der handlungsorientierten Fehlertaxonomie (Zapf et al., 1989) 31 kritische Ereignisse (93.94 %) zu insgesamt 10 einzigartigen Nutzungsproblemen konsolidiert wurden (d.h. weniger strikte Anwendung von IntuiBeat-F). Von diesen 10 Nutzungsproblemen

wurden bei einem Nutzungsproblem ein Bewegungsfehler (10 %) als Ursache klassifiziert, welches sich aus drei der 33 kritischen Ereignissen (9.10 %) ableiten ließ. Es handelt sich dementsprechend nur bei neun der 10 Nutzungsprobleme (90 %) auch um reale Nutzungsprobleme, die zur Sicherstellung einer intuitiven Benutzung beseitigt werden müssen. Diese realen Probleme wurden aus 28 der 33 kritischen Ereignisse (84.85 %) abgeleitet. Die restlichen zwei von 33 kritischen Ereignissen (6.06 %) stellten lediglich (zeitliche) Ineffizienzen im Sinne der handlungsorientierten Fehlertaxonomie dar und bildeten dadurch ebenfalls keine realen Nutzungsprobleme ab. Bei der Einteilung der Nutzungsprobleme in die Fehlerkategorien der handlungsorientierten Fehlertaxonomie von Zapf et al. (1989) konnte ein Cohens κ von .63 festgestellt werden, was laut (Landis & Koch, 1977) als eine beachtliche Übereinstimmung interpretiert werden kann und damit auf einem ähnlich hohem Niveau wie in den Vorgängerexperimenten lag ($\kappa_{Experiment4} = .78$; $\kappa_{Experiment5} = .80$).

Zusammenfassend kann festgehalten werden, dass die Trefferrate von IntuiBeat-F bei wenig strikter Anwendung bezogen auf alle Nutzungsprobleme damit bei 93.94 % und die Fehlalarmrate bei 6.06 % lag. Betrachtet man nur die aus kritischen Ereignissen abgeleiteten realen Nutzungsprobleme, reduziert sich die Trefferrate bei wenig strikter Anwendung entsprechend auf 84.85 % und die Fehlalarmrate steigt auf 15.15 % (siehe Tabelle 6.6).

Explorative Datenanalyse von IntuiBeat-F (strikte Anwendung)

Betrachtet man ausschließlich die mithilfe von Rhythmus-Peaks erkannten kritischen Ereignisse (d.h. strikte Anwendung von IntuiBeat-F), wurden insgesamt 39 Rhythmus-Peaks generiert, wobei sich entsprechend 30 Rhythmus-Peaks (76.92 %) auf 10 Nutzungsprobleme bzw. 28 Rhythmus-Peaks (71.80 %) auf neun reale Nutzungsprobleme verteilten. Zwei der 39 Rhythmus-Peaks (5.13 %) wiesen damit auf einen Bewegungsfehler hin. Die restlichen neun der 39 Rhythmus-Peaks (23.08 %) gingen dementsprechend nicht in Nutzungsprobleme ein und führten auch nicht zur Entdeckung (zeitlicher) Ineffizienzen im Sinne der handlungsorientierten Fehlertaxonomie. Betrachtet man ausschließlich die durch jeweils einen Rhythmus-Peak erkannten 30 von 33 (90.91 %) insgesamt mit IntuiBeat-F gefundenen kritischen Ereignisse, gingen hierbei alle mithilfe von Rhythmus-Peaks identifizierten kritischen Ereignisse (100 %) in insgesamt 10 der mit IntuiBeat-F gefundenen Nutzungsprobleme (100 %) ein. Beschränkt man sich nur auf die mit IntuiBeat-F abgeleiteten realen Nutzungsprobleme, konnten 28 der 30 durch Rhythmus-Peaks identifizierten, auf reale Nutzungsprobleme hinweisenden, kritischen Ereignisse (93.33 %) alle neun realen Nutzungsprobleme (100 %) identifizieren. Schließlich gingen zwei von 30 (6.67 %) insgesamt durch Rhythmus-Peaks identifizierten kritischen Ereignissen in ein Nutzungsproblem ein, welches einen Bewegungsfehler als Ursache hatte. Dementsprechend lag die Trefferrate bei strikter Anwendung von IntuiBeat-F bei der Ableitung aller Nutzungsprobleme bei 100 % und die Fehlalarmrate bei 0 %, was einer Verbesserung von rund 6 % gegenüber der wenig strikten Anwendung entsprach. Betrachtet man nur die durch strikte Anwendung von IntuiBeat-F abgeleiteten realen Nutzungsprobleme kann eine Trefferrate von 93.33 % und eine Fehlalarmrate von 6.67 % festgestellt werden, was sogar einer Steigerung von über 8 % gegenüber der wenig strikten Anwendung entspricht (siehe Tabelle 6.6).

Zusammenfassend kann also festgehalten werden, dass vom Evaluator lediglich 76.92 % aller 39 Rhythmus-Peaks für die Ableitung von Nutzungsproblemen berücksichtigt und damit 23.08 % der Rhythmus-Peaks ignoriert wurden. Betrachtet man lediglich reale Nutzungsprobleme, wurden entsprechend nur 71.80 % aller 39 Rhythmus-Peaks für die Ableitung von realen Nutzungsproblemen genutzt und damit sogar 28.20% der Rhythmus-Peaks vernachlässigt. In diesem Zusammenhang wurden lediglich 90.91 % von den gesamten 33 kritischen Ereignissen mithilfe der 39 Rhythmus-Peaks entdeckt, was bezogen auf die 31 auf Nutzungsprobleme hinweisenden kritischen Ereignisse 96.77 % (30 Rhythmus-Peaks) und bezogen auf die 28, auf reale Nutzungsprobleme hinweisenden, kritischen Ereignisse 100 % (28 Rhythmus-Peaks) darstellte. Dementsprechend wurden nur 9.10 % aller kritischen Ereignisse unabhängig von Rhythmus-Peaks protokolliert. Bezogen auf alle Nutzungsprobleme wurden damit 3.23 % und bezogen auf alle realen Nutzungsprobleme 0 % der kritischen Ereignisse unabhängig festgehalten. Aus diesem Grund konnten bei der ausschließlichen Berücksichtigung von Rhythmus-Peaks bei IntuiBeat-F (d.h. strikte Anwendung von IntuiBeat-F) keine Nutzungsprobleme (d.h. 100 % auffindbar) und kein reales Nutzungsproblem weniger (d.h. 100 % auffindbar) bei gleichzeitiger Steigerung der Trefferrate bzw. Senkung der Fehlalarmrate beobachtet werden (siehe Tabelle 6.6).

Explorative Datenanalyse des Nutzertests mit retrospektivem Think-Aloud-Protokoll

Mithilfe des Nutzertests mit retrospektivem Think-Aloud-Protokoll konnte 34 kritische Ereignisse erkannt werden, wovon unter Berücksichtigung der in der Abbildung 6.2 in Teilabschnitt 6.1.3 beschriebenen Arbeitsschritte zur Ableitung von Nutzungsproblemen (Burmester, 2016) und der handlungsorientierten Fehlertaxonomie (Zapf et al., 1989) 30 kritische Ereignisse (88.24 %) zu insgesamt sechs einzigartigen Nutzungsproblemen konsolidiert wurden, wovon 21 kritische Ereignisse (70 %) zu fünf Nutzungsproblemen führten, die einen Bewegungsfehler als Ursache hatten (83.33 %) und bei denen es sich damit nicht um reale Nutzungsprobleme handelte. Die restlichen vier der 34 kritischen Ereignisse (11.76 %) stellten lediglich (zeitliche) Ineffizienzen im Sinne der handlungsorientierten Fehlertaxonomie und damit auch keine echten Nutzungsproblem im Bezug auf intuitive Benutzung dar. Dementsprechend ließ sich ein reales Nutzungsproblem aus neun der 34 kritischen Ereignisse ableiten (26.47 %).

Zusammenfassend kann festgehalten werden, dass die Trefferrate bezogen auf alle aus kritischen Ereignissen abgeleiteten Nutzungsprobleme geringer als bei IntuiBeat-F bei 88.24 % und die Fehlalarmrate höher als bei IntuiBeat-F bei 11.74 % lag. Betrachtet man nur die aus kritischen Ereignissen abgeleiteten realen Nutzungsprobleme reduziert sich die Trefferrate entsprechend auf 26.47 % und die Fehlalarmrate steigt auf 73.53 % (siehe Tabelle 6.6).

Qualitative Ursachenanalyse mit beiden Methoden gefundener Nutzungsprobleme

Beurteilt man die mithilfe von IntuiBeat-F (d.h. weniger strikte Anwendung) und dem Nutzertest mit retrospektivem Think-Aloud-Protokoll abgeleiteten Nutzungsprobleme mit der handlungsorientierten Fehlertaxonomie qualitativ, stellt man bei Berücksichtigung aller kritischen Ereignisse folgende Einteilung fest (siehe Abbildung 6.13). Im Falle von IntuiBeat-F konnten drei Gewohnheitsfehler (60 %), ein Erkennensfehler (20 %) und ein Bewegungsfehler (20 %) als Ursachen für die fünf Nutzungsprobleme klassifiziert werden. Damit konnten die Ursachen hauptsächlich der perzeptiv-begrifflichen Ebene zugeschrieben werden. Bezüglich des Nutzertests mit retrospektivem Think-Aloud-Protokoll konnten ein Wissensfehler (16.57 %) und fünf Bewegungsfehler (83.33 %) als Ursachen für die sechs einzigartigen Nutzungsprobleme klassifiziert werden. Die Ursache lag damit überwiegend auf der sensomotorischen Regulationsebene. Die Ursachen für die von beiden Methoden unter Berücksichtigung aller kritischen Ereignisse abgeleiteten fünf Nutzungsprobleme waren vier Gewohnheitsfehler (80 %) und ein Wissensfehler (20 %). Die von beiden Methoden gefundenen Ursachen teilen sich dementsprechend mit 80 % auf die perzeptiv-begriffliche Regulationsebene und mit 20 % auf die Regulationsgrundlage auf. Berücksichtigt man bei dieser Beurteilung bei IntuiBeat-F ausschließlich die mithilfe des Algorithmus abgeleiteten Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F), ergibt sich eine identische Einteilung (siehe Abbildung 6.13).

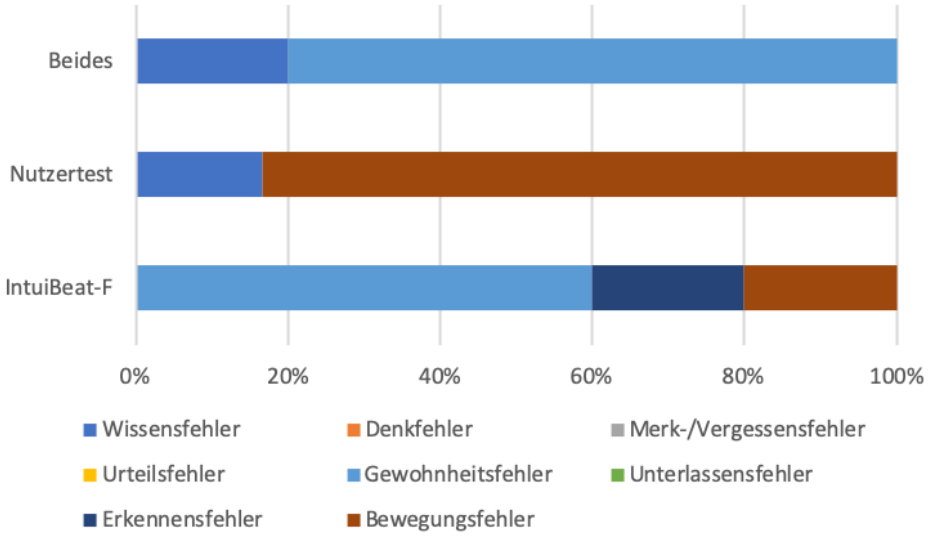


Abbildung 6.13. Ursachenklassifikation mithilfe der handlungsorientierten Fehlertaxonomie (Zapf, Brodbeck, & Prümper, 1989) der mit IntuiBeat-F und einem Nutzertest mit retrospektivem Think-Aloud-Protokoll abgeleiteten Nutzungsprobleme im Rahmen des sechsten Experiments.

6.3.5 Diskussion

Wie bereits in den beiden Experimenten zuvor, wurde im vorliegenden sechsten Experiment die wissenschaftliche Güte von IntuiBeat-F bezüglich der Gütekriterien *Gründlichkeit* und *Gültigkeit*, sowie der *zeitlichen Anwendungseffizienz*, als für das Projekt 3D-GUIde wichtiges Kriterium der praktischen Güte, beim Vergleich mit einem als Quasi-Außenkriterium fungierenden Nutzertest mit retrospektivem Think-Aloud-Protokoll überprüft. Als Untersuchungsgegenstand fungierte die CUI-Software *Tinkercad*, bei der es sich auf Basis von Experteneinschätzungen und Nutzereinschätzungen (anhand eines TFQ) um eine eher weniger intuitiv benutzbare Software handelt und es entsprechend zu vielen, intuitive Benutzung beeinträchtigender Nutzungsproblemen (d.h. reale Nutzungsprobleme) kommen sollte. Im Vergleich zu den beiden Vorgängerexperimenten kam eine zusätzliche Analysesoftware während des retrospektiven Interviews zum Einsatz, um den Evaluator die Zuordnung von Rhythmus-Peaks zu Videostellen zu erleichtern und dadurch die wissenschaftliche Güte, sowie zeitliche Anwendungseffizienz von IntuiBeat-F zu verbessern.

Anhand der statistischen Tests wurde zunächst überprüft, ob sich ein Unterschied bezüglich der Gründlichkeit zu Gunsten von IntuiBeat-F feststellen lässt (Hypothese H1). Anschließend wurde geprüft, ob auch bezüglich der Gültigkeit ein Unterschied zu Gunsten von IntuiBeat-F erkennbar ist (Hypothese H2). Abschließend wurde untersucht, inwiefern IntuiBeat-F eine höhere zeitliche Anwendungseffizienz als der Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der formativen Evaluation intuitiver Benutzung des CUI *Tinkercad* aufweisen kann (Hypothese H3). Im folgenden Verlauf werden die Hypothesen bezüglich der Gründlichkeit, der Gültigkeit und der zeitlichen Anwendungseffizienz unter Berücksichtigung der festgestellten Ergebnisse diskutiert, nachdem im Vorfeld auf eine mögliche Konfundierung durch die *Vorerfahrung bei der Nutzung von CUIs* eingegangen wurde.

6.3.5.1 Überprüfung einer möglichen Konfundierung durch die Vorerfahrung bei der Nutzung von CUIs

Die Vorerfahrung bei der Nutzung von CUIs unterschied sich erwartungsgemäß nicht signifikant zwischen den beiden Ausprägungen der unabhängigen Variable *Art der formativen Evaluationsmethode*. Jedoch war die Teststärke aufgrund des geringen beobachteten Effekts und der kleinen Stichprobe gering. Wie bereits zuvor beim vierten Experiment (siehe Abschnitt 6.1), bei dem ebenfalls ein weniger intuitiv benutzbares CUI formativ bezüglich intuitiver Benutzung evaluiert wurde, könnte eine mögliche Erklärung für dieses Ergebnis darin bestehen, dass bei den als Stichprobe genutzten Studierenden der Medienkommunikation und der Mensch-Computer-Systeme keine großen Unterschiede bezüglich der Vorerfahrung zu erwarten sind (d.h. aufgrund der nicht vorhandenen Anforderungen von CUI-Kenntnissen und der allgemeinen Studieninhalte, der in der Stichprobe berücksichtigten Studiengänge, sind bezüglich der Vorerfahrung keine größeren Schwankungen zu erwarten und diese ähnlich gering ausgeprägt).

Dementsprechend ist die praktische Bedeutsamkeit der in diesem Zusammenhang ermittelten Effekte bzw. der Erkenntnisgewinn aufgrund der verwendeten Stichprobe eher als unbedeutend einzustufen (siehe Döring & Bortz, 2016), was sich auch in den deskriptiven

Unterschieden erkennen lässt. Da darüber hinaus bei der Kontrollvariable *Vorerfahrung bei der Nutzung von CUIs* keine signifikanten Korrelationen bezüglich der Gründlichkeit, der Gültigkeit und der zeitlichen Anwendungseffizienz bei beiden formativen Evaluationsmethoden vorlagen ($p > .05$), lag mit hoher Wahrscheinlichkeit keine Konfundierung durch eine unterschiedliche Vorerfahrung bei der Nutzung von CUIs vor.

6.3.5.2 Überprüfung der Gründlichkeit

Wie erwartet, zeigte sich im Gegensatz zu den beiden Vorgängerexperimenten die höhere Gründlichkeit von IntuiBeat-F gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der Evaluation einer weniger intuitiv benutzbaren Software mit zusätzlicher Analysesoftware inferenzstatistisch, sowohl bei Berücksichtigung aller Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F), als auch bei ausschließlicher Berücksichtigung mithilfe von Rhythmus-Peaks entdeckter Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F). Der Einsatz einer zusätzlichen Analysesoftware im sechsten Experiment führte damit im Vergleich zu den beiden Vorgängerexperimenten dazu, dass der Evaluator lediglich drei kritische Ereignisse unabhängig von Rhythmus-Peaks identifizierte und von ihm auch über 70 % der von IntuiBeat-F angezeigten Rhythmus-Peaks beim retrospektivem Interview letztendlich für die Identifikation von realen Nutzungsproblemen genutzt wurden. Dementsprechend wurden über alle Versuchspersonen hinweg auch keine realen Nutzungsprobleme mithilfe von Rhythmus-Peaks und damit identifizierten kritischen Ereignissen übersehen (siehe Teilabschnitt 6.3.4.6). Die von IntuiBeat-F erreichte Gründlichkeit lag demzufolge sowohl bei wenig strikter als auch strikter Anwendung von IntuiBeat-F bei 27 % und damit signifikant höher als die Gründlichkeit des Nutzertests mit retrospektivem Think-Aloud-Protokoll, die lediglich bei 18 % lag (siehe Teilabschnitt 6.3.4.3). Damit ähneln die beiden Ergebnisse des sechsten Experiments von IntuiBeat-F den Ergebnissen aus den Vorgängerexperimenten (d.h. Gründlichkeit von 32 % beim vierten und 29 % beim fünften Experiment), wobei jedoch bei IntuiBeat-F alle Nutzungsprobleme vom Evaluator berücksichtigt wurden (d.h. weniger strikte Anwendung von IntuiBeat-F) und nicht nur die ausschließlich mithilfe von Rhythmus-Peaks identifizierbaren Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F), da sich bei letzteren kein signifikanter Unterschied im Vergleich zum Nutzertest mit retrospektivem Think-Aloud-Protokoll zeigte (siehe Teilabschnitte 6.1.5 und 6.2.5).

Wo der Evaluator in den Vorgängerexperimenten kritische Ereignisse ohne die zusätzliche Unterstützung durch eine Analysesoftware entdecken musste und dadurch nicht alle Rhythmus-Peaks im Rahmen des retrospektiven Interviews berücksichtigen konnte, unterstützte ihn die neue Analysesoftware bei der Berücksichtigung aller Rhythmus-Peaks im sechsten Experiment mit hoher Wahrscheinlichkeit, weswegen konzeptuell kein Unterschied zwischen der strikten und der weniger strikten Anwendung von IntuiBeat-F zu beobachten war (siehe Teilabschnitt 6.3.4.3). Damit liegen die Ergebnisse des sechsten Experiments bezüglich der Gründlichkeit ebenfalls im Rahmen des laut Koutsabasis et al. (2007) bei formativen Evaluationsmethoden für Gebrauchstauglichkeit (d.h. nicht speziell intuitive Benutzung) üblichen Wertebereichs von circa 20 bis 40 %, wenn als Evaluatoren Studierende und damit Novizen genutzt werden, die zwar über eine große Kenntnis des

Untersuchungsgegenstands verfügen, es ihnen jedoch noch an Erfahrung bei der Durchführung eines Nutzertests fehlt. Darüber hinaus ist hier anzumerken, dass die hohen Werte bezüglich der Gründlichkeit in den Vorgängerexperimenten mit hoher Wahrscheinlichkeit dadurch zustande gekommen sind, dass der Evaluator unabhängig von Rhythmus-Peaks reale Nutzungsprobleme abgeleitet hat. Wie bereits im Rahmen der Diskussionen bezüglich der Gründlichkeit in den beiden letzten Experimenten angesprochen (siehe Teilabschnitte 6.1.5 und 6.2.5), war dies sehr wahrscheinlich, da nur ein geringer Anteil an Rhythmus-Peaks überhaupt berücksichtigt wurde (d.h. rund 40 % der Rhythmus-Peaks waren unberücksichtigt), ein hoher Anteil von kritischen Ereignissen, die auf reale Nutzungsprobleme hinweisen konnten, unberücksichtigt blieb (d.h. rund 60 % blieben unberücksichtigt) und sich in Folge dessen auch die Anzahl an gefundenen realen Nutzungsproblemen bei strikter Anwendung von IntuiBeat-F reduzierte (d.h. rund 20 % der realen Nutzungsprobleme blieben unberücksichtigt).

Betrachtet man nur die ausschließlich auf Rhythmus-Peaks rückführbaren Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F), ergibt sich gegenüber den Vorgängerexperimenten jedoch ein anderes Bild (d.h. Gründlichkeit bei ausschließlicher Berücksichtigung von Rhythmus-Peaks lag bei 19 % im vierten und bei 14 % im fünften Experiment) und im Vergleich der Ergebnisse eine Verbesserung von über 10 % bei strikter Anwendung von IntuiBeat-F durch die zusätzlich eingesetzte Analysesoftware im sechsten Experiment. Da alle Nutzungsprobleme auch mithilfe von Rhythmus-Peaks abgeleitet werden konnten und damit die Gründlichkeit sowohl bei strikter als auch weniger strikter Anwendung von IntuiBeat-F konzeptuell gleich war, konnte logischerweise durch die Einführung der neuen Analysesoftware auch keine Steigerung der Gründlichkeit durch die strikte Anwendung im Vergleich zur wenig strikten Anwendung von IntuiBeat-F erreicht werden (siehe Teilabschnitt 6.3.4.3). Dieser nicht vorhandene Unterschied zeigt jedoch im Vergleich zu den Ergebnissen aus den beiden Vorgängerexperimenten deutlich, dass mithilfe der neuen Analysesoftware die Ableitung realer Nutzungsprobleme von der Willkür des Evaluators und dessen Expertise entkoppelt wurde, was durch eine statistisch signifikant höhere Gründlichkeit bei der strikten Anwendung von IntuiBeat-F im Vergleich zu einem Nutzertest mit retrospektivem Think-Aloud-Protokoll nachgewiesen werden konnte. Diese starke Entkoppelung von der Expertise des Evaluators kann aber auch nachteilig sein, da man durch den Einsatz der Analysesoftware Gefahr laufen kann, die identifizierten kritischen Ereignisse nicht zu hinterfragen, was der nicht vorhandene Unterschied zwischen der wenig strikten und strikten Anwendung von IntuiBeat-F vermuten lässt (siehe Teilabschnitt 6.3.4.3), und dadurch insbesondere das Wissen von erfahrenen Evaluatoren nicht ausgeschöpft werden kann. Die Software würde den Evaluator sozusagen immer nur dabei unterstützen, ein Minimum an Gründlichkeit zu erreichen, könnte jedoch auch die aufgrund der Expertise des Evaluators maximal mögliche Gründlichkeit künstlich beschränken.

Setzt man die Ergebnisse des sechsten Experiments zusätzlich ins Verhältnis mit anderer Forschung zu formativen Evaluationsmethoden aus dem Forschungsgebiet der Gebrauchstauglichkeit (d.h. nicht speziell intuitive Benutzung), in der laut der Meta-Analysen von Virzi (1992) und Nielsen und Landauer (1993) mit fünf Probanden bereits 85 % aller realen Nutzungsprobleme auffindbar sind, sind die im sechsten Experiment erzielten Ergebnisse von IntuiBeat-F bezüglich Gründlichkeit auf jeden Fall mehr als zufriedenstellend. Virzi (1992) und Nielsen und Landauer (1993) kamen unabhängig voneinander auf die folgende kumulative binomische Wahrscheinlichkeitsformel (siehe Formel 6.3) zur Abschätzung der

Gründlichkeit (d.h. s) bzw. Auftretenswahrscheinlichkeit von realen Nutzungsproblemen bei Berücksichtigung von n Versuchspersonen, wobei Virzi (1992) diese Formel mithilfe von Monte-Carlo-Simulationen und Nielsen und Landauer (1993) die Formel mithilfe eines Poisson-Modells auf Basis vorhandener empirischer Daten ableiteten (siehe Formel 6.3).

$$\text{Anteil gefundener realer Nutzungsprobleme} = 1 - (1 - s)^n \tag{6.3}$$

Um die im Rahmen des sechsten Experiments ermittelten Werte von IntuiBeat-F und des Nutzertests mit retrospektivem Think-Aloud-Protokoll bezüglich Gründlichkeit mit der Literatur vergleichen zu können, wurden diese Werte in die obige Formel eingesetzt. Bei IntuitBeat-F und einer Gründlichkeit von 27 % würde es demnach sechs Versuchspersonen benötigen um 85 % der Probleme zu finden, wohingegen beim Nutzertest mit retrospektivem Think-Aloud-Protokoll aufgrund der geringeren Gründlichkeit von 18 % bereits 10 Versuchspersonen notwendig wären. Man könnte an dieser Stelle anmerken, dass beide Methoden bereits über der „magischen Grenze“ von fünf Versuchspersonen liegen und demnach ihre Gründlichkeit bereits als zufriedenstellend beurteilt werden kann. Jedoch konnten aktuellere, etwas konservativere Meta-Analysen, wie beispielsweise die von Spool und Schroeder (2001) oder die von Faulkner (2003), zeigen, dass mit fünf Versuchspersonen oftmals eine Gründlichkeit von lediglich 35 % bzw. lediglich 55 % in der Realität machbar ist.

Auch Akers et al. (2012), die mithilfe ihrer automatisch generierten Backtracking-Operationen lediglich eine Gründlichkeit von 4.2 % erreichten, benötigten basierend auf der Formel bereits 10 Versuchspersonen um überhaupt die 35 % von Spool und Schroeder (2001) zu erreichen. Die mit IntuiBeat-F und der zusätzlichen Analysesoftware erreichte Gründlichkeit ist demnach nicht nur zufriedenstellend, sondern vielmehr beeindruckend. Aufgrund der Tatsache, dass diese „Fünf-Personen-Regel“ oftmals aufgrund ihrer Generalität missverstanden wird und in der Praxis formative Evaluationsstudien oft nicht mit einer angemessenen Stichprobengröße durchgeführt werden, schätzen einige Forscher wie Lewis (2014) und Borsci et al. (2013) den Bedarf an Versuchspersonen mithilfe von verschiedenen statistischen Modellen (z.B. Monte-Carlo, Good-Turing, siehe Borsci et al., 2013) auf Basis von zwei Versuchspersonen ab und passen ihr statistisches Modell entsprechend nach zwei weiteren Versuchspersonen an. Auch unter Berücksichtigung dieses Vorgehens, bei dem acht Versuchspersonen zu 80 % aller realen Nutzungsprobleme führen, kann die Gründlichkeit von IntuiBeat-F als hoch interpretiert werden. Für eine vertiefte und aktuellere Auseinandersetzung mit dieser Thematik sei an dieser Stelle auf Caine (2016) und Borsci et al. (2013) verwiesen.

Nichtsdestotrotz soll aus Gründen der Replizierbarkeit in einem weiteren, abschließenden Folgeexperiment geklärt werden, ob die Gründlichkeit von IntuiBeat-F in Kombination mit einer zusätzlichen Analysesoftware auch bei einer stärker intuitiv benutzbare Software, bei der erwartungsgemäß mit weniger realen Nutzungsproblemen und damit auch bei einem gewöhnlichen Nutzertest grundsätzlich mit einer höheren Gründlichkeit gerechnet werden kann, im Vergleich zu einem Nutzertest mit retrospektivem Think-Aloud-Protokoll ebenfalls signifikant höher liegt. Unabhängig davon können die Ergebnisse bezüglich der Gründlichkeit von IntuiBeat-F auch bei Einsatz der zusätzlichen Analysesoftware bei der

Evaluation einer weniger intuitiv benutzbaren Software als mehr als zufriedenstellend interpretiert und zusammenfassend die in diesem Zusammenhang aufgestellten Hypothesen H1.A und H1.B entsprechend als bestätigt angesehen werden (siehe Teilabschnitt 6.7).

6.3.5.3 Überprüfung der Gültigkeit

Wie erwartet, zeigte sich die höhere Gültigkeit von IntuiBeat-F gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der Evaluation einer weniger intuitiv benutzbaren Software mit zusätzlicher Analysesoftware inferenzstatistisch, sowohl bei Berücksichtigung aller Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F), als auch bei Berücksichtigung nur durch den Algorithmus entdeckter Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F). In beiden Fällen konnte bei IntuiBeat-F bei Berücksichtigung aller in dieser Bedingung identifizierten Nutzungsprobleme eine signifikant höhere Gültigkeit von rund 90 % (d.h. weniger strikt: 88 %, strikt: 90 %) gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll beobachtet werden, dessen Gültigkeit lediglich bei 51 % (d.h. deskriptiver Unterschied von 37 % bzw. 39 %) lag (siehe Teilabschnitt 6.3.4.4), was im Vergleich zu anderen formativen Evaluationsmethoden innerhalb der Usability-Forschung (d.h. Evaluationsmethoden die Gebrauchstauglichkeit formativ messen, aber nicht speziell intuitive Benutzung) ähnlich hoch bzw. niedrig (d.h. zwischen 74 und 90 % bei empirischen Nutzertests) liegt (siehe Koutsabasis et al., 2007). Im Vergleich zu den beiden Vorgängerexperimenten, bei denen keine zusätzliche Analysesoftware zum Einsatz kam, lag die Gültigkeit von IntuiBeat-F jedoch bei Berücksichtigung aller Nutzungsprobleme (d.h. wenig strikte Anwendung von IntuiBeat-F) mit 88 % deskriptiv zwischen den ermittelten Ergebnisse der Vorgängerexperimente (d.h. Gültigkeit von 92 % im vierten und Validität von 79 % im fünften Experiment), stellt aber unter Bezug auf die Arbeit von Koutsabasis et al. (2007) immer noch ein mehr als zufriedenstellendes Ergebnis dar.

Betrachtet man die Gültigkeit des sechsten Experiments ausschließlich unter Berücksichtigung der durch Rhythmus-Peaks abgeleiteten Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F), steigt diese von 88 % lediglich um 2 % auf 90 %, was darin begründet sein kann, dass im Vergleich zu den beiden Vorgängerexperimenten, bei denen weniger als 50 % der kritischen Ereignisse berücksichtigt wurden, im sechsten Experiment nur knapp 3 % aller kritischen Ereignisse unberücksichtigt blieben. Betrachtet man bei strikter Anwendung von IntuiBeat-F nur reale Nutzungsprobleme, blieben im sechsten Experiment sogar keine auf solche Nutzungsprobleme hinweisenden kritischen Ereignisse unberücksichtigt. Im Vergleich zu den Vorgängerexperimenten wurde im Zuge dessen auch der Anteil nicht berücksichtigter Rhythmus-Peaks reduziert. Der Anteil an realen Nutzungsproblemen an allen Nutzungsproblemen veränderte sich dadurch im Vergleich zu den Vorgängerexperimenten weniger stark (d.h. Experiment 4: Steigerung der Gültigkeit um 4 %; Experiment 5: Steigerung der Gültigkeit um 18 %), da alle Nutzungsproblemen und alle realen Nutzungsproblem auch mithilfe von Rhythmus-Peaks erkennbar waren und damit der Anteil entgangener Nutzungsprobleme bei Null lag (siehe Teilabschnitt 6.3.4.6). Die 2 % kommen daher nur aufgrund individueller Unterschiede zustande, da die Gültigkeit ja auf Versuchspersonenebene berechnet wurde und es hier dennoch zu interindividueller Varianz gekommen sein kann. Da im sechsten Experiment eine neue Analysesoftware für

die einfachere Zuordnung der Rhythmus-Peaks eingesetzt wurde, sprechen die genannten Zahlen und die geringen Unterschiede bezüglich der Gültigkeit zwischen der Berücksichtigung aller Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) und der Berücksichtigung ausschließlich mit Rhythmus-Peaks assoziierter Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F) für die Effektivität der neuen Analysesoftware. Nichtsdestotrotz waren die 10 gefundenen Nutzungsprobleme und die neun realen gefundenen Nutzungsproblemen im Vergleich (siehe Teilabschnitt 6.3.4.6) zu der Anzahl, die in den beiden Vorgängerexperimenten ermittelt wurde, etwas geringer (d.h. Experiment 4: 60 gesamt, 56 real; Experiment 5: 50 gesamt, 47 real).

Eine mögliche Erklärung kann die Tatsache liefern, dass wie bereits im Rahmen der Diskussion der Gründlichkeit angesprochen, generell eine hohe Variabilität bei der Anzahl gefundener Nutzungsprobleme besteht (z.B. Borsci et al., 2013; Caine, 2016; Faulkner, 2003; Lewis, 2014) und damit auch der eingesetzte Evaluator eine große Rolle spielt (z.B. Hertzum & Jacobsen, 2001; Hertzum et al., 2014; Hornbæk & Frøkjær, 2008; Jacobsen et al., 1998). Obwohl bei der Auswahl der Evaluatoren bewusst darauf geachtet wurde, in allen Experimenten Evaluatoren mit gleichem Kenntnisstand zu verwenden und den Evaluator pro Experiment in beiden Versuchsbedingungen zur besseren Kontrolle des Evaluatoreffekts konstant zu halten, kann nicht ausgeschlossen werden, dass der Evaluator einen ungewollten Einfluss hatte. Aufgrund der Tatsache, dass sich jedoch beim Nutzertest mit retrospektivem Think-Aloud-Protokoll im Vergleich zu IntuiBeat-F im sechsten Experiment eine noch viel geringere Anzahl entdeckter Nutzungsprobleme zeigte (d.h. 6 gesamt, 1 real, siehe Teilabschnitt 6.3.4.6) war der Einfluss in beiden Versuchsbedingungen mit hoher Wahrscheinlichkeit zumindest ähnlich hoch. IntuiBeat-F und die zusätzliche Analysesoftware konnten den Evaluator trotzdem bei der Identifikation von (realen) Nutzungsproblemen unterstützen, was ein erstes Indiz dafür sein kann, dass IntuiBeat-F auch Evaluatoren, die generell wenig (reale) Nutzungsprobleme entdecken können (d.h. Novizen mit geringer Vorerfahrung), stark unterstützen kann, da mithilfe von IntuiBeat-F im Vergleich zum Nutzertest mit retrospektivem Think-Aloud-Protokoll auf diese Weise immerhin vier Nutzungsprobleme insgesamt und sogar acht reale Nutzungsprobleme mehr gefunden werden konnten (siehe Teilabschnitt 6.3.4.6).

Weitere mögliche Erklärungen könnten auch darin bestehen, dass womöglich im Vergleich zu den vorigen Experimenten die Aufgaben trotz Experteneinschätzung doch zu einfach waren (d.h. Grund war der gewählte Untersuchungsgegenstand selbst) oder es bei der Zuordnung der Rhythmus-Peaks zu den entsprechenden Videostellen trotz Analysesoftware zu Fehlern gekommen ist. Ebenso wären Fehler auch bei der softwareunabhängigen Einordnung der Nutzungsprobleme in die handlungsorientierte Fehlertaxonomie von Zapf et al. (1989) möglich, obwohl aufgrund der hohen Beurteilerübereinstimmung eher weniger damit zu rechnen ist. Obwohl die Gültigkeit von IntuiBeat-F beim sechsten Experiment stets (d.h. bei strikter und weniger strikter Anwendung) statistisch signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll lag (siehe Teilabschnitt 6.3.4.4), soll aus Gründen der Replizierbarkeit in einem abschließenden Folgeexperiment geklärt werden, ob die Gültigkeit von IntuiBeat-F auch unter Verwendung einer zusätzlichen Analysesoftware bei einer stärker intuitiv benutzbare Software, bei der erwartungsgemäß mit weniger Nutzungsproblemen (d.h. real und nicht real) und damit auch bei einem gewöhnlichen Nutzertest grundsätzlich mit einer höheren Gültigkeit gerechnet werden kann, im Vergleich zu einem Nutzertest mit retrospektivem Think-Aloud-Protokoll ebenfalls signi-

fikant höher liegt. Unabhängig davon können die Ergebnisse bezüglich der Gültigkeit von IntuiBeat-F auch bei Einsatz der zusätzlichen Analysesoftware bei der Evaluation einer weniger intuitiv benutzbaren Software als mehr als zufriedenstellend interpretiert und die in diesem Zusammenhang aufgestellten Hypothesen H2.A und H2.B entsprechend als bestätigt angesehen werden (siehe Tabelle 6.7).

6.3.5.4 Überprüfung der zeitlichen Anwendungseffizienz

Wie erwartet, zeigte sich die signifikant höhere zeitliche Anwendungseffizienz von IntuiBeat-F gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll auch bei der Evaluation einer weniger intuitiv benutzbaren Software unter Einsatz einer zusätzlichen Analysesoftware. IntuiBeat-F nahm sogar circa 40 % weniger Zeit für die Anwendung in Anspruch (siehe Teilabschnitt 6.3.4.5), was aufgrund des sehr geringen Anteils nicht berücksichtigter, auf reale Nutzungsprobleme hinweisender, Rhythmus-Peaks (d.h. lediglich rund 20 %) und der vollständigen Berücksichtigung auf reale Nutzungsprobleme hinweisender kritischer Ereignisse auch zu erwarten war (siehe Teilabschnitt 6.3.4.6).

Im Vergleich zu den beiden Vorgängerexperimenten (d.h. circa 20 % beim vierten Experiment und circa 25 % beim fünften Experiment gesteigerte zeitliche Anwendungseffizienz im Vergleich zum Nutzertest mit retrospektivem Think-Aloud-Protokoll), bei denen die Zuordnung der Rhythmus-Peaks zu den entsprechenden Videostellen noch vollständig manuell durchgeführt werden musste, konnte durch den Einsatz der neuen Analysesoftware eine Verbesserung der zeitlichen Anwendungseffizienz von rund 15 % erreicht werden (siehe Teilabschnitt 6.3.4.5). Die in diesem Zusammenhang aufgestellte Hypothese H3 kann dementsprechend bei der Evaluation einer weniger intuitiv benutzbaren Software als bestätigt angesehen werden, wenn zusätzlich eine Analysesoftware während des retrospektiven Interviews zum Einsatz kommt (siehe Teilabschnitt 6.7). Wie bereits im Rahmen der Diskussionen zur Gründlichkeit und Gültigkeit angesprochen, soll dieses Ergebnis noch in einem abschließenden Folgeexperiment anhand einer konzeptuellen Replikation mit einer stärker intuitiv benutzbaren Software und der Analysesoftware verifiziert werden, da bei einer stärker intuitiv benutzbaren Software im Allgemeinen auch bei einem gewöhnlichen Nutzertest aufgrund weniger zu berücksichtigender Nutzungsprobleme mit einer höheren zeitlichen Anwendungseffizienz gerechnet werden kann.

6.3.6 Schlussfolgerung

Zusammenfassend kann festgehalten werden, dass die wissenschaftliche Güte von IntuiBeat-F als formative Evaluationsmethode für intuitive Benutzung hinsichtlich des Gütekriteriums *Gültigkeit* auch bei der Untersuchung einer weniger intuitiv gestalteten Software bei Verwendung einer zusätzlichen Analysesoftware empirisch, sowohl bei strikter als auch weniger strikter Anwendung, bestätigt werden konnte (siehe Tabelle 6.7). Darüber hinaus konnte IntuiBeat-F aufgrund seiner im Vergleich zum Nutzertest mit retrospektivem Think-Aloud-Protokoll höheren zeitlichen Anwendungseffizienz auch ein für das Projekt 3D-GUIde wichtiger Teilaspekt praktischer Güte empirisch zugesprochen werden (siehe Tabelle 6.7). Konfundierungen durch Unterschiede in der Vorerfahrung bei der Nutzung von CUIs können mit hoher Wahrscheinlichkeit ausgeschlossen werden. Die Effekt- und

Teststärken lagen dabei überwiegend im oberen Bereich. Im Gegensatz zu den beiden Vorgängerexperimenten, in denen keine zusätzliche Analysesoftware zum Einsatz kam und der Evaluator Rhythmus-Peaks mit hoher Wahrscheinlichkeit aufgrund der komplizierten manuellen Zuordnung von Rhythmus-Peaks zu kritischen Ereignissen nicht berücksichtigt hatte, konnte auch die wissenschaftliche Güte von IntuiBeat-F bezüglich des Gütekriteriums der Gründlichkeit sowohl bei strikter als auch weniger strikter Anwendung von IntuiBeat-F bestätigt werden (siehe Tabelle 6.7).

Tabelle 6.7. *Übersicht der mithilfe des sechsten Experiments bestätigten Hypothesen im Zuge der Meta-Evaluation von IntuiBeat-F (KEs: kritische Ereignisse; RP-KEs: mit Rhythmus-Peaks assoziierte kritische Ereignisse).*

Hypothese	Experiment 6
(H1) Überprüfung der Gründlichkeit:	
- (A) Alle KEs	✓
- (B) Nur RP-KEs	✓
(H2) Überprüfung der Gültigkeit:	
- (A) Alle KEs	✓
- (B) Nur RP-KEs	✓
(H3) Überprüfung der zeitlichen Anwendungseffizienz	✓

Da im Vergleich zu den beiden Vorgängerexperimenten die Gründlichkeit von IntuiBeat-F zwar bei strikter Anwendung größer war, die Werte bezüglich Gültigkeit sich jedoch zwischen Ergebnissen der Vorgängerexperimente befanden und die zeitliche Anwendungseffizienz gegenüber den Vorgängerexperimenten doppelt so hoch war, soll aus Gründen der Replizierbarkeit noch ein Folgeexperiment mit demselben Evaluator durchgeführt werden. Im Rahmen dieses Experiments soll abschließend verifiziert werden, ob die wissenschaftliche Güte und zeitliche Anwendungseffizienz von IntuiBeat-F unter Verwendung einer zusätzlichen Analysesoftware bei einer stärker intuitiv benutzbaren Software im Vergleich zu einem Nutzertest mit retrospektivem Think-Aloud-Protokoll ebenfalls bestätigt werden kann, wo es erwartungsgemäß zu weniger realen Nutzungsproblemen kommt und damit auch bei einem gewöhnlichen Nutzertest mit einer höheren Gründlichkeit, Gültigkeit und zeitlichen Anwendungseffizienz im Allgemeinen gerechnet werden kann.

6.4 Experiment 7

Das in diesem Abschnitt vorgestellte siebte und letzte Experiment verfolgte, wie die vorigen drei Experimente zuvor, das Ziel zu überprüfen, inwiefern IntuiBeat-F wissenschaftliche Güte bezüglich der Hauptgütekriterien *Gründlichkeit* und *Gültigkeit*, sowie *zeitliche Anwendungseffizienz* attestiert werden kann. Hierzu kam es zum Vergleich mit einem als Quasi-Außenkriterium fungierenden Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der Nutzung der Kochwebseite *Chefkoch.de* (Chefkoch GmbH, 2017), weswegen das siebte Experiment im Vergleich zum vorigen Experiment eine konzeptuelle Replikation darstellt (siehe Tabelle 6.7), die ebenfalls bei der Durchführung des retrospektiven Interviews die neu entwickelte Analysesoftware berücksichtigte und versuchte die Ergebnisse

des vorigen Experiments in einer anderen Domäne bzw. bei der Nutzung einer stärker intuitiv benutzbaren Software zu replizieren (siehe Diskussion des vorigen Experiments in Teilabschnitt 6.3.5). Das siebte Experiment adressierte dementsprechend auch die zweite Forschungsfrage und den formativen Aspekt der dritten Forschungsfrage dieser Arbeit.

6.4.1 Überprüfung der Gründlichkeit und Gültigkeit

Aufgrund der Tatsache, dass es sich beim siebten Experiment um eine konzeptuelle Replikation des vierten Experiments unter Berücksichtigung einer neuen Domäne und einer zusätzlichen Analysesoftware für das retrospektive Interview handelte, wurden auch die gleichen Hypothesen wie bei den Vorgängerexperimenten aufgestellt, weswegen an dieser Stelle auf den entsprechenden Teilabschnitt 6.1.1 des vierten Experiments verwiesen werden soll.

6.4.2 Überprüfung der zeitlichen Anwendungseffizienz

Auch bezüglich der zeitlichen Anwendungseffizienz wurde die Hypothese aus den Vorgängerexperimenten übernommen, weswegen an dieser Stelle auf den entsprechenden Teilabschnitt 6.1.2 des vierten Experiments verwiesen werden soll.

6.4.3 Methode

6.4.3.1 Teilnehmer

Für das siebte Experiment wurden 27 Versuchspersonen über das Probandensystem des Instituts für Mensch-Computer-Medien an der Universität Würzburg rekrutiert. Aufgrund der Tatsache, dass bei einer Person die Rhythmusaufzeichnung nicht vollständig (d.h. Rhythmusaufzeichnung bei einer Teilaufgabe vergessen) und bei zwei Personen die Videoaufzeichnung unvollständig war (d.h. Aufnahme stürzte ab), mussten diese Datensätze von der Datenauswertung ausgeschlossen werden. Demzufolge konnten für die Meta-Evaluation von IntuiBeat-F 24 Versuchspersonen berücksichtigt werden, welche alle rechtsfüßig (d.h. der rechte Fuß stellte den dominanten Fuß dar) waren. Die Versuchspersonen setzten sich dabei aus 17 Frauen und sieben Männern zusammen. Das Durchschnittsalter betrug 21.29 Jahre ($SD = 3.06$). Es handelte sich bei allen Teilnehmern um Studierende der Julius-Maximilians-Universität Würzburg, wobei fünf Personen Mensch-Computer-Systeme (20.83 %) und 19 Personen Medienkommunikation (79.17 %) studierten. Alle Versuchsteilnehmer wurden über das Probanden-System des Instituts Mensch-Computer-Medien über eine gesonderte Mail darauf hingewiesen, für den Versuch flache Sportschuhe zu tragen, um eine möglichst problemlose Rhythmuseingabe über das USB-Fußpedal zu ermöglichen. Für die Teilnahme an der Untersuchung bekam jede Versuchsperson eine halbe Versuchspersonenstunde gutgeschrieben. Die mit einem TFQ gemessene Vorerfahrung der Versuchspersonen bezüglich der Nutzung von Kochwebseiten betrug im Durchschnitt 1.53 ($SD = .86$) bei einem Maximum von 6 und lag damit im höheren Bereich. Alle Versuchspersonen besaßen damit eine mittelmäßige Vorerfahrung mit Kochwebseiten. Alle Versuchspersonen gaben an, am Experiment freiwillig teilzunehmen.

6.4.3.2 Versuchsdesign

Für die Beantwortung der zweiten Forschungsfrage und des formativen Aspekts der dritten Forschungsfrage wurde das gleiche Experimentaldesign wie zuvor beim vierten Experiment genutzt. Für genauere Informationen bezüglich der unabhängigen und abhängigen Variablen wird daher auf den entsprechenden Absatz 6.1.3.1 des vierten Experiments verwiesen.

6.4.3.3 Versuchsmaterialien und Maße

Die im siebten Experiment verwendeten Versuchsmaterialien und Maßen unterschieden sich nur in einigen Punkten von den im sechsten Experiment verwendeten Versuchsmaterialien und Maßen, weswegen in diesem Teilabschnitt auch nur auf diese eingegangen wird und für die übernommenen Versuchsmaterialien und Maße auf den entsprechenden Absatz 6.3.3.3 des sechsten Experiments verwiesen werden soll.

Untersuchungsgegenstand der formativen Evaluation: Chefkoch.de

Als Untersuchungsgegenstand für die Meta-Evaluation von IntuiBeat-F kam die Kochwebseite *Chefkoch.de* der Chefkoch GmbH (2017) als stärker intuitiv benutzbare Software auf Basis einer qualitativen Experteneinschätzung ($N_{Experte}$= 5; Vorgehen: siehe entsprechenden Absatz innerhalb des Teilabschnitts 5.1.4.3) im Rahmen des siebten Experiments zum Einsatz (siehe Abbildung 6.14). Wie bereits beim fünften Experiment wurde explizit kein CUI als Untersuchungsgegenstand verwendet, da mit dem siebten Experiment die Robustheit der Ergebnisse des vorherigen Experiments, bei dem beim retrospektiven Interview eine zusätzliche Analysesoftware zur Unterstützung zum Einsatz kam, anhand einer konzeptuellen Replikation unter Berücksichtigung dieser zusätzlichen Unterstützung mit einem Untersuchungsgegenstand aus einer anderen Domäne bzw. einer stärker intuitiv benutzbaren Software überprüft werden sollte.

Es wurde sich als Untersuchungsgegenstand für eine Kochwebseite entschieden, da unter Berücksichtigung der für das Experiment zur Verfügung stehenden Studierendenstichprobe zu erwarten ist, dass Chefkoch.de genauso wie die Hotelbuchungswebseite *HolidayCheck.de* zuvor, von Studierenden überwiegend als stärker intuitiv benutzbare Softwareanwendung eingestuft wird (d.h. stärker intuitiv benutzbare Softwareanwendungen sollten logischerweise auch zu weniger Nutzungsproblemen führen), da so gut wie jede Person eine derartige Website schon einmal genutzt hat. Aufgrund der Tatsache, dass es sich bei Chefkoch.de wie HolidayCheck.de ebenfalls um eine Webseite handelte und dementsprechend eine ähnliche Dynamik wie bei CUIs vorliegt (siehe entsprechender Absatz des Teilabschnitts 6.2.3.3), sollte mithilfe von Chefkoch.de die wissenschaftliche Güte von IntuiBeat-F anhand der Gütekriterien *Gründlichkeit* und *Gültigkeit* im Vergleich zum vorigen Experiment nicht nur unter Berücksichtigung der zusätzlichen Unterstützung beim retrospektiven Interview durch die Analysesoftware und einer anderen Domäne, sondern auch unter Berücksichtigung einer stärker intuitiv benutzbaren Softwareanwendung (vergleiche Tinkercad als weniger intuitiv benutzbare Softwareanwendung mit vielen Nutzungsproblemen) und damit

auch bei einer Anwendung mit höchstwahrscheinlich weniger Nutzungsproblemen nachgewiesen werden.

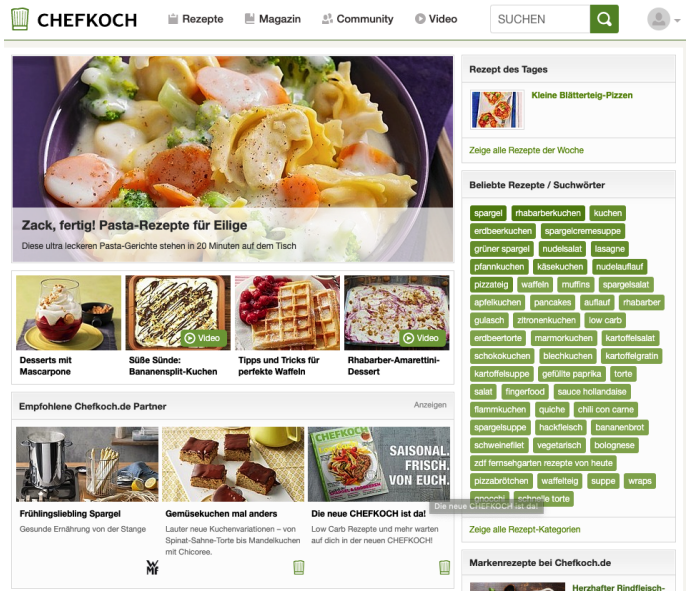


Abbildung 6.14. Die im Rahmen des siebten Experiments als Untersuchungsgegenstand verwendete Kochwebseite *Chefkoch.de* der Chefkoch GmbH (2017).

Es wurden auf Basis eigener Nutzungserfahrungen mit Kochwebseiten, fünf experimentelle Aufgaben gewählt (siehe Anhang B.7.1), die den Funktionsumfang von Chefkoch.de so gut wie möglich repräsentieren sollten. Sie wurden von den Versuchspersonen in einer festen Reihenfolge bearbeitet. Alle Aufgaben wurden dabei in ein fiktives „Kochszenario“ eingebettet, in dem es um das gemeinsame Kochen von Freunden ging. Hierbei mussten die Versuchspersonen zunächst einen neuen Einkaufszettel auf Chefkoch.de anlegen, da sie für ein Risotto noch Parmesan und Weißwein einkaufen sollten (erste Aufgabe). Diese Einkaufsliste mussten die Versuchspersonen im Anschluss als „Risotto“ benennen (zweite Aufgabe) und dieser Einkaufsliste anschließend den Parmesan, sowie den Weißwein hinzufügen (dritte Aufgabe). Daraufhin mussten die Versuchspersonen die fertige Einkaufsliste in „Einkaufsliste Risotto“ umbenennen (vierte Aufgabe). Abschließend mussten die Versuchspersonen noch nach einem Rezept für ein Dessert suchen, was zusammen mit dem Risotto den Gästen serviert werden sollte. Aufgrund der Tatsache, dass die Freunde noch Eier und Butter daheim hatten, und ein Gast keine Zartbitterschokolade mag, musste das Rezept diese Einschränkungen berücksichtigen (fünfte Aufgabe).

Vorerfahrung bei der Nutzung von Kochwebseiten

Die Vorerfahrung der Versuchspersonen bezüglich der Kochwebseite *Chefkoch.de* wurde mit einem hierfür konstruierten TFQ als Kontrollvariable erhoben (KV_Vorerfahrung), welcher wie in den Experimenten zuvor, schriftlich administriert wurde. Als Items für diesen Fragebogen wurden lediglich „Chefkoch.de“ und „Anderes Webportal zum Thema Kochen?“ genutzt, wobei letzteres Item eine Wildcard darstellte und die Angabe einer beliebigen bekannten Kochwebseite erlaubte. Es wurde bei der Konstruktion des TFQ auf die

Nennung konkreter Softwareanwendungen verzichtet, da es heutzutage ein Überangebot an Kochwebseiten gibt, was eine repräsentative Auswahl erschwert. Da wahrscheinlich jeder der als Stichprobe fungierenden Studierenden der Medienkommunikation und Mensch-Computer-Systeme schon einmal ein Rezept auf einer solchen Seite gesucht hat, erschien die Wildcard ausreichen, um die damit verbundene Expertise erfassen zu können.

Wie bei den anderen konstruierten TFQs, beurteilten die Versuchspersonen diese beiden Items dann bezüglich ihrer Nutzungshäufigkeit (Wertebereich: 0 - 6) und des genutzten Funktionsumfangs (Wertebereich: 0 - 4). Im Anschluss wurde dann für jede Dimension (d.h. Nutzungshäufigkeit und Funktionsumfang) ein Mittelwert gebildet und daraufhin ein Gesamtmittelwert über beide Mittelwerte als Operationalisierung der *Vorerfahrung bei der Nutzung von Kochwebseiten* berechnet (Wertebereich: 0 - 5). Die Entscheidung für eine Mittelwertbildung anstelle einer Summenberechnung wurde im Rahmen des ersten Experiments ausführlich begründet, weswegen an dieser Stelle lediglich darauf verwiesen wird (siehe Absatz „Vorerfahrung bei der Nutzung von CUIs“ des Teilabschnitts 5.1.4.3). Da kein einheitlicher TFQ existiert und dieser jeweils für die jeweiligen getesteten Softwareanwendungen erstellt werden musste, ist der in diesem Experiment genutzte TFQ vollständig in Anhang B.7.2 dieser Arbeit zu finden.

6.4.3.4 Versuchsdurchführung

Die Versuchsdurchführung des siebten Experiments war im Vergleich zur Versuchsdurchführung der drei Vorgängerexperimente mit der Ausnahme identisch, dass die Versuchspersonen mit der Kochwebseite *Chefkoch.de* arbeiteten, weswegen an dieser Stelle für die Beschreibung der Versuchsdurchführung auf den entsprechenden Teilabschnitt 6.1.3.4 des vierten Experiments verwiesen werden soll.

6.4.3.5 Statistische Auswertung

Die statistische Auswertung des siebten Experiments war identisch mit der statistischen Auswertung der drei Vorgängerexperimente, weswegen für genauere Informationen auf den entsprechenden Teilabschnitt 6.1.3.5 des vierten Experiments verwiesen werden soll. Da die *Vorerfahrung bei der Nutzung von Kochwebseiten*, die als Kontrollvariable erhoben wurde, einen ungewollten Einfluss auf die Gründlichkeit und Gültigkeit der verglichenen Methoden haben könnte, musste eine derartige Konfundierung vor der Überprüfung der Gründlichkeit von IntuiBeat-F ausgeschlossen werden. Hierzu wurden t-Tests für unabhängige Stichproben bezüglich der Vorerfahrung, sowie Pearson-Produkt-Moment-Korrelationen zwischen der Vorerfahrung und den abhängigen Variablen innerhalb der beiden Ausprägungen der unabhängigen Variable gerechnet (siehe Teilabschnitt 6.4.4.2). Da in beiden Fällen keine signifikanten Ergebnisse festgestellt werden konnten (d.h. kein ungewollter Einfluss der Vorerfahrung auf Gründlichkeit und Gültigkeit), wurde von einer univariaten Kovarianzanalyse (ANCOVA) abgesehen (Döring & Bortz, 2016; Field, 2017) und sich stattdessen für eine Datenauswertung mithilfe von t-Tests für unabhängige Stichproben ohne Berücksichtigung der Vorerfahrung bei der Nutzung von Kochwebseiten als Kovariate entschieden (siehe Abschnitt 6.4.4). Die Überprüfung der statistischen Voraussetzungen

der Pearson-Produkt-Moment-Korrelationen und der t-Tests werden im folgenden Ergebnisteil berichtet (siehe Teilabschnitt 6.4.4.1).

6.4.4 Ergebnisse

Im folgenden Abschnitt werden die Ergebnisse bezüglich der in den Teilabschnitten 6.4.1 und 6.4.2 beschriebenen Hypothesen deskriptiv und inferenzstatistisch berichtet. Vor der eigentlichen Datenanalyse wird zunächst auf die Überprüfung der statistischen Voraussetzungen eingegangen. Dabei wurde bei allen abhängigen Variablen und Kontrollvariablen ein metrisches Skalenniveau angenommen.

6.4.4.1 Überprüfung der statistischen Voraussetzungen

Überprüfung von Ausreißern

Das Vorgehen bei der Ausreißeranalyse des siebten Experiments war identisch mit dem Vorgehen des vierten Experiments, weswegen für genauere Informationen auf den entsprechenden Teilabschnitt 6.1.4.1 des vierten Experiments verwiesen werden soll. Es mussten auf diese Weise bei keiner der analysierten abhängigen Variablen und Kontrollvariablen Werte ausgeschlossen werden.

Überprüfung der Voraussetzung der Normalverteilung

Die univariate Normalverteilung der abhängigen Variablen und Kontrollvariablen wurde für jede Ausprägung der unabhängigen Variable mittels Kolmogorov-Smirnov-Tests ($p \geq .05$, siehe Field, 2017) und Sichtprüfung anhand eines Q-Q-Diagramms geprüft. Dabei konnten auf Basis dieser beiden Kriterien bei der AV_Gründlichkeit_WenigerStrikt und bei der AV_Gründlichkeit_Strikt im Rahmen der Überprüfung der Gründlichkeit (Hypothese H1), sowie bei der AV_Gültigkeit_WenigerStrikt und bei der AV_Gültigkeit_Strikt im Rahmen der Überprüfung der Gültigkeit (Hypothese H2) keine Normalverteilung in beiden Gruppen festgestellt werden. Aufgrund der Tatsache, dass ein ungepaarter t-Test bei etwa gleich großen Gruppen robust gegenüber Verletzungen der Normalverteilungsannahme ist (Glass et al., 1972; Pagano, 2006; Salkind, 2010; Wilcox, 2011), wurde sich für eine eindeutige Interpretation und Vergleichbarkeit der Ergebnisse (d.h. Mittelwerte statt Medianen) der verschiedenen Experimente gegen Transformationen und entsprechende nonparametrische Verfahren zur Überprüfung der Hypothesen (d.h. H1 und H2) entschieden.

Überprüfung der Voraussetzung der Homoskedastizität

Zur Überprüfung der Homoskedastizität zwischen den Ausprägungen der unabhängigen Variable kamen Levene-Tests zum Einsatz (siehe Field, 2017), welche im Rahmen der Überprüfung der Gründlichkeit (Hypothese H1), der Überprüfung der Gültigkeit (Hypothese H2) und der Überprüfung der zeitlichen Anwendungseffizienz (Hypothese H3)

gerechnet und bei allen abhängigen Variablen und Kontrollvariablen ($p \geq .05$, siehe Field, 2017) außer bei AV_Gründlichkeit_WenigerStrikt und AV_Gründlichkeit_Strikt Varianzhomogenität bestätigen konnten. Die Freiheitsgrade für den dazugehörigen t-Test im Rahmen der Überprüfung der Gründlichkeit (Hypothese H1) wurden dementsprechend korrigiert.

6.4.4.2 Überprüfung einer möglichen Konfundierung durch Vorerfahrung bei der Nutzung von Kochwebseiten

Die Vorerfahrung bei der Nutzung von Kochwebseiten (Wertebereich: 0 - 5) unterschied sich nicht signifikant zwischen den beiden Ausprägungen der unabhängigen Variable (Nutzertest mit retrospektivem Think-Aloud-Protokoll: $M = 1.52$; $SD = .91$; IntuiBeat-F: $M = 1.54$, $SD = .85$), $t(22) = -.06$, $p = .954$, d = .02. Laut J. Cohen (1988) kann dieser Effekt als klein interpretiert werden ($d \leq .5$). Eine konservative post-hoc Analyse der Teststärke bezüglich KV_Vorerfahrung mithilfe von G*Power (Faul et al., 2009) mit $df = 22$ und einer angenommenen geringen Effektstärke (d = .02) ergab lediglich eine geringe Teststärke (1 - $\beta = .06$) im Zuge der Auswertung der Vorerfahrung bei der Nutzung von Kochwebseiten. Um jedoch eine ausreichend große Power (1 - $\beta \geq .80$) beim festgestellten Effekt erzielen zu können, wären pro Gruppe 30428 Versuchspersonen nötig gewesen, was aufgrund des straffen Zeitplans im Anwenderprojekt, des dortigen Fokus auf qualitative Ergebnisse, des Verständnisses der Meta-Evaluation von IntuiBeat-F als Nebenprodukt und der personellen Einschränkungen im Projekt 3D-GUIde nicht möglich gewesen wäre. Auch bei Annahme einer großen Effektstärke (d = .8) hätten immerhin noch 26 Versuchspersonen pro Gruppe getestet werden müssen, was im Hinblick auf die Einschränkungen durch das Anwenderprojekt nicht notwendig erschien.

Da die erhobene Vorerfahrung bei der Nutzung von CUIs dennoch einen ungewollten Einfluss auf den Vergleich der beiden Ausprägungen der unabhängigen Variable haben könnte, wurde für die Kontrollvariable mithilfe von Pearson-Produkt-Moment-Korrelationen sichergestellt, dass kein Zusammenhang zwischen dieser und den abhängigen Variablen innerhalb der beiden Ausprägungen der unabhängigen Variable besteht ($p > .05$). Die Kontrollvariable KV_Vorerfahrung wurde dementsprechend nicht als Kovariate in der statistischen Auswertung berücksichtigt, da kein linearer Zusammenhang zwischen dieser und den abhängigen Variablen bestand (siehe Döring & Bortz, 2016; Field, 2017).

6.4.4.3 Überprüfung der Gründlichkeit

Wie erwartet, lag die Gründlichkeit (%) von IntuiBeat-F ($M = .16$, $SD = .08$) signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll ($M = .10$, $SD = .04$), wenn alle mit IntuiBeat-F abgeleiteten Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) berücksichtigt wurden (H1.A), $t(16.36) = -2.35$, $p = .031$, d = .95. Es war laut J. Cohen (1988) ein großer Effekt feststellbar (d $\geq$ .8). Aufgrund der Tatsache, dass alle mit IntuiBeat-F abgeleiteten Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) auch mithilfe des Algorithmus abgeleitet werden konnten (d.h. strikte Anwendung von IntuiBeat-F), konnte auch bei der strikten Anwendung ein

statistisch identisches Ergebnis wie bei der weniger strikten Anwendung erzielt werden (H1.B).

6.4.4.4 Überprüfung der Gültigkeit

Wie erwartet, lag die Gültigkeit (%) von IntuiBeat-F (M = .90, SD = .15) signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll (M = .66, SD = .18), wenn alle mit IntuiBeat-F abgeleiteten Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) berücksichtigt wurden (H2.A), $t(22)$ = - 3.60, p = .002, d = 1.45. Es war laut J. Cohen (1988) ein großer Effekt feststellbar (d $\geq$ .8). Auch wenn ausschließlich die auf Basis des Algorithmus von IntuiBeat-F abgeleitete Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F) berücksichtigt werden (H2.B), lag die Gültigkeit von IntuiBeat-F (M = .95, SD = .12), wie erwartet, höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll (M = .66, SD = .18), $t(22)$ = - 4.68, $p < .001$, d = 1.90. Es war hier laut J. Cohen (1988) ebenfalls ein großer Effekt feststellbar (d $\geq$ .8).

6.4.4.5 Überprüfung der zeitlichen Anwendungseffizienz

Wie erwartet, lag die zeitliche Anwendungseffizienz (s) von IntuiBeat-F in Form einer geringeren durchschnittlichen Problemidentifikationszeit für reale Nutzungsprobleme (M = 352.50, SD = 154.69) signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll (M = 603.89, SD = 200.73), $t(22)$ = 3.44, p = .002, d = 1.40. Es war laut J. Cohen (1988) ein großer Effekt feststellbar (d $\geq$ .8).

6.4.4.6 Explorative Datenanalyse

Wie bereits in Teilabschnitt 6.1.3.3 des vierten Experiments angesprochen, wurden zusätzlich eine Reihe von Metriken für eine explorative Datenanalyse berechnet, mit deren Hilfe verstanden werden kann, warum es zu keinem Unterschied bezüglich Gründlichkeit (siehe Teilabschnitt 6.4.4.3) und zu einem Unterschied bezüglich Gültigkeit (siehe Teilabschnitt 6.4.4.4) bei der strikten und weniger strikten Anwendung von IntuiBeat-F gekommen ist. Im Folgenden sollen diese Metriken nun berichtet werden.

Tabelle 6.8. *Metriken zur explorativen Datenanalyse von IntuiBeat-F und des Nutzertests mit retrospektivem Think-Aloud-Protokoll im Rahmen des siebten Experiments (NPs: Nutzungsprobleme; KEs: kritische Ereignisse; RP-KEs: mit Rhythmus-Peaks assoziierte kritische Ereignisse; Alle NPs aus KEs ≈ weniger strikte Anwendung von IntuiBeat-F; Nur NPs aus RP-KEs ≈ strikte Anwendung von IntuiBeat-F).*

Metrik	Experiment 7	
	IntuiBeat-F	Nutzertest
Trefferrate NPs (%):		
- Alle NPs aus KEs	94.12	92.31
- Nur NPs aus RP-KEs	100	-
Trefferrate reale NPs (%):		
- Alle NPs aus KEs	82.35	57.69
- Nur NPs aus RP-KEs	90.32	-
Fehlalarmrate NPs (%):		
- Alle NPs aus KEs	5.88	7.69
- Nur NPs aus RP-KEs	0	-
Fehlalarmrate reale NPs (%):		
- Alle NPs aus KEs	17.65	42.31
- Nur NPs aus RP-KEs	9.68	-
Anzahl NPs:		
- Alle NPs aus KEs	15	8
- Nur NPs aus RP-KEs	14	-
Anzahl realer NPs:		
- Alle NPs aus KEs	13	3
- Nur NPs aus KEs mit RPs	13	-
Anteil realer NPs an NPs (%):		
- Alle NPs aus KEs	86.67	37.50
- Nur NPs aus RP-KEs	92.86	-

Explorative Datenanalyse von IntuiBeat-F (weniger strikte Anwendung)

Mithilfe von IntuiBeat-F konnten über alle Versuchspersonen hinweg 34 kritische Ereignisse erkannt werden, wovon unter Berücksichtigung der in der Abbildung 6.2 in Teilabschnitt 6.1.3 beschriebenen Arbeitsschritte zur Ableitung von Nutzungsproblemen (Burmester, 2016) und der handlungsorientierten Fehlertaxonomie (Zapf et al., 1989) 32 kritische Ereignisse (94.12 %) zu insgesamt 15 einzigartigen Nutzungsproblemen konsolidiert wurden (d.h. weniger strikte Anwendung von IntuiBeat-F). Von diesen 15 Nutzungsproblemen

wurden bei zwei Nutzungsproblemen ein Bewegungsfehler als Ursache klassifiziert (13.33 %), welche sich aus vier der 34 kritischen Ereignissen (11.77 %) ableiten ließen. Es handelt sich dementsprechend nur bei 13 der 15 Nutzungsprobleme (86.67 %) auch um reale Nutzungsprobleme, die zur Sicherstellung einer intuitiven Benutzung beseitigt werden müssen. Diese realen Probleme wurden aus 28 der 34 kritischen Ereignisse (82.35 %) abgeleitet. Die restlichen zwei der 34 kritischen Ereignisse (5.88 %) stellten lediglich (zeitliche) Ineffizienzen im Sinne der handlungsorientierten Fehlertaxonomie dar und bildeten dadurch ebenfalls keine realen Nutzungsprobleme ab. Bei der Einteilung der Nutzungsprobleme in die Fehlerkategorien der handlungsorientierten Fehlertaxonomie von Zapf et al. (1989) konnte ein Cohens κ von .70 festgestellt werden, was laut Landis und Koch (1977) als eine beachtliche Übereinstimmung interpretiert werden kann und damit auf einem ähnlich hohem Niveau wie die Vorgängerexperimente lag ($\kappa_{Experiment4} = .78$; $\kappa_{Experiment5} = .80$; $\kappa_{Experiment6} = .63$).

Zusammenfassend kann festgehalten werden, dass die Trefferrate von IntuiBeat-F bei wenig strikter Anwendung bezogen auf alle Nutzungsprobleme damit bei 94.12 % und die Fehlalarmrate bei 5.88 % lag. Betrachtet man nur die aus kritischen Ereignissen abgeleiteten realen Nutzungsprobleme, reduziert sich die Trefferrate bei wenig strikter Anwendung entsprechend auf 82.35 % und die Fehlalarmrate steigt auf 17.65 % (siehe Tabelle 6.8).

Explorative Datenanalyse von IntuiBeat-F (strikte Anwendung)

Betrachtet man ausschließlich die mithilfe von Rhythmus-Peaks erkannten kritischen Ereignisse (d.h. strikte Anwendung von IntuiBeat-F), wurden insgesamt 46 Rhythmus-Peaks generiert, wobei sich entsprechend 31 Rhythmus-Peaks (67.39 %) auf 14 Nutzungsprobleme bzw. 28 Rhythmus-Peaks (60.87 %) auf 13 reale Nutzungsprobleme verteilten. Drei der 46 Rhythmus-Peaks (6.52 %) bildeten dementsprechend einen Bewegungsfehler ab. Die restlichen 15 der 46 Rhythmus-Peaks (32.61 %) gingen somit nicht in Nutzungsprobleme ein und führten auch nicht zur Entdeckung (zeitlicher) Ineffizienzen im Sinne der handlungsorientierten Fehlertaxonomie. Betrachtet man ausschließlich die durch jeweils einen Rhythmus-Peak erkannten 31 von 34 (91.18 %) insgesamt mit IntuiBeat-F gefundenen kritischen Ereignisse, gingen hierbei alle mithilfe von Rhythmus-Peaks identifizierten kritischen Ereignisse (100 %) in insgesamt 14 der 15 insgesamt mit IntuiBeat-F gefundenen Nutzungsprobleme (93.33 %) ein. Beschränkt man sich nur auf die mit IntuiBeat-F abgeleiteten realen Nutzungsprobleme, konnten 28 der 31 insgesamt durch Rhythmus-Peaks identifizierten, auf reale Nutzungsprobleme hinweisenden, kritischen Ereignisse (90.32 %) alle 13 realen Nutzungsprobleme (100 %) identifizieren. Schließlich gingen drei von 31 (9.68 %) insgesamt durch Rhythmus-Peaks identifizierte kritische Ereignisse in ein Nutzungsproblem ein, welches einen Bewegungsfehler als Ursache hatte. Dementsprechend lag die Trefferrate bei strikter Anwendung von IntuiBeat-F bei der Ableitung aller Nutzungsprobleme bei 100 % und die Fehlalarmrate bei 0 %, was einer Verbesserung von rund 6 % gegenüber der wenig strikten Anwendung entspricht. Betrachtet man nur die durch strikte Anwendung von IntuiBeat-F abgeleiteten realen Nutzungsprobleme kann ebenfalls eine Trefferrate von 90.32 % und eine Fehlalarmrate von 9.68 % festgestellt werden, was sogar einer Steigerung von 8 % gegenüber der wenig strikten Anwendung entsprach (siehe Tabelle 6.8).

Zusammenfassend kann also festgehalten werden, dass vom Evaluator lediglich 67.39 % aller 46 Rhythmus-Peaks für die Ableitung von Nutzungsproblemen berücksichtigt und damit 32.61 % der Rhythmus-Peaks ignoriert wurden. Betrachtet man lediglich reale Nutzungsprobleme, wurden entsprechend nur 60.87 % aller 46 Rhythmus-Peaks für die Ableitung von realen Nutzungsproblemen genutzt und damit sogar 39.13 % der Rhythmus-Peaks vernachlässigt. In diesem Zusammenhang wurden lediglich 91.18 % von den gesamten 34 kritischen Ereignissen mithilfe der 31 Rhythmus-Peaks entdeckt, was bezogen auf die 32 auf Nutzungsprobleme hinweisenden kritischen Ereignisse 96.88 % (31 Rhythmus-Peaks) und bezogen auf die 28, auf reale Nutzungsprobleme hinweisenden, kritischen Ereignisse 100 % (28 Rhythmus-Peaks) darstellte. Dementsprechend wurden nur 8.82 % aller kritischen Ereignisse unabhängig von Rhythmus-Peaks protokolliert. Bezogen auf alle Nutzungsprobleme wurden damit 3.13 % und bezogen auf alle realen Nutzungsprobleme 0 % der kritischen Ereignisse unabhängig festgehalten. Aus diesem Grund konnten bei der ausschließlichen Berücksichtigung von Rhythmus-Peaks bei IntuiBeat-F (d.h. strikte Anwendung von IntuiBeat-F) lediglich ein Nutzungsproblem (d.h. noch 93.33 % auffindbar) und kein reales Nutzungsproblem weniger (d.h. 100 % auffindbar) bei gleichzeitiger Steigerung der Trefferrate bzw. Senkung der Fehlalarmrate beobachtet werden (siehe Tabelle 6.8).

Explorative Datenanalyse des Nutzertests mit retrospektivem Think-Aloud-Protokoll

Mithilfe des Nutzertests mit retrospektivem Thinking-Aloud-Protokoll konnten 26 kritische Ereignisse erkannt werden, wovon unter Berücksichtigung der in der Abbildung 6.2 in Teilabschnitt 6.1.3 beschriebenen Arbeitsschritte zur Ableitung von Nutzungsproblemen (Burmester, 2016) und der handlungsorientierten Fehlertaxonomie (Zapf et al., 1989) 24 kritische Ereignisse (92.31 %) zu insgesamt acht einzigartigen Nutzungsproblemen konsolidiert wurden, wovon neun der 26 kritischen Ereignisse (34.62 %) zu fünf von acht Nutzungsproblemen führten, die einen Bewegungsfehler als Ursache hatten (20 %) und es sich damit nicht um reale Nutzungsprobleme handelte. Die restlichen zwei der 26 kritischen Ereignisse (7.69 %) stellten lediglich (zeitliche) Ineffizienzen im Sinne der handlungsorientierten Fehlertaxonomie und damit auch keine echten Nutzungsproblem im Bezug auf intuitive Benutzung dar. Dementsprechend ließ sich drei reale Nutzungsprobleme aus 15 der 26 kritischen Ereignisse ableiten (57.69 %).

Zusammenfassend kann festgehalten werden, dass die Trefferrate bezogen auf alle aus kritischen Ereignissen abgeleiteten Nutzungsproblemen geringer als bei IntuiBeat-F bei 92.31 % und die Fehlalarmrate höher als bei IntuiBeat-F bei 7.69 % lag. Betrachtet man nur die aus kritischen Ereignissen abgeleiteten realen Nutzungsprobleme reduziert sich die Trefferrate entsprechend auf 57.69 % und die Fehlalarmrate steigt auf 42.31 % (siehe Tabelle 6.8).

Qualitative Ursachenanalyse mit beiden Methoden gefundener Nutzungsprobleme

Beurteilt man die mithilfe von IntuiBeat-F (d.h. weniger strikte Anwendung von IntuiBeat-F) und dem Nutzertest mit retrospektivem Think-Aloud-Protokoll abgeleiteten Nutzungsprobleme mit der handlungsorientierten Fehlertaxonomie qualitativ, stellt man bei Berücksichtigung aller kritischen Ereignisse folgende Einteilung fest (siehe Abbildung 6.15). Im Falle von IntuiBeat-F konnten ein Gewohnheitsfehler (14.285 %), ein Unterlassensfehler (14.285 %), drei Erkennensfehler (42.86 %), ein Urteilsfehler (14.285 %) und ein Bewegungsfehler (14.285 %) als Ursachen für die sieben Nutzungsprobleme klassifiziert werden. Damit konnte die Ursachen hauptsächlich der perzeptiv-begrifflichen Ebene zugeschrieben werden. Bezüglich des Nutzertests mit retrospektivem Think-Aloud-Protokoll konnten ein Gewohnheitsfehler (12.50 %), ein Erkennensfehler (12.50 %), ein Urteilsfehler (12.50 %) und fünf Bewegungsfehler (62.50 %) als Ursachen für die acht einzigartigen Nutzungsprobleme klassifiziert werden. Die Ursache lag damit überwiegend auf der sensomotorischen Regulationsebene. Die Ursachen für die von beiden Methoden unter Berücksichtigung aller kritischen Ereignisse abgeleiteten acht Nutzungsprobleme waren fünf Gewohnheitsfehler (62.50 %), zwei Urteilsfehler (25 %) und ein Bewegungsfehler (12.50 %). Die von beiden Methoden gefundenen Ursachen teilen sich dementsprechend mit 62.50 % auf die perzeptiv-begriffliche Regulationsebene, mit 25 % auf die intellektuelle Regulationsebene und mit 12.50 % auf die sensomotorische Regulationsebene auf. Berücksichtigt man bei dieser Beurteilung bei IntuiBeat-F ausschließlich die mithilfe des Algorithmus abgeleiteten Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F), ergibt sich eine identische Einteilung (siehe Abbildung 6.15).

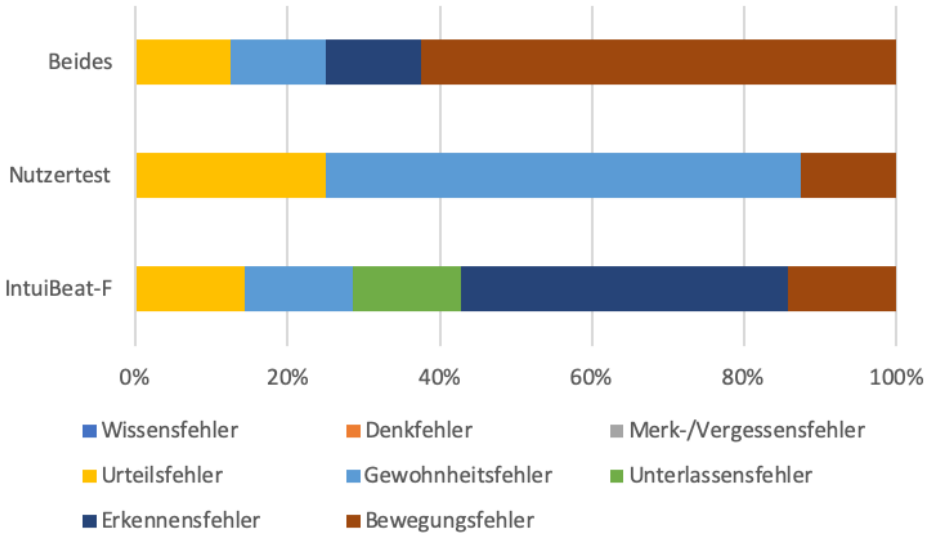


Abbildung 6.15. Ursachenklassifikation mithilfe der handlungsorientierten Fehlertaxonomie (Zapf, Brodbeck, & Prümper, 1989) der mit IntuiBeat-F und einem Nutzertest mit retrospektivem Think-Aloud-Protokoll abgeleiteten Nutzungsprobleme im Rahmen des siebten Experiments.

6.4.5 Diskussion

Wie bereits im Experiment zuvor, wurde im vorliegenden siebten Experiment die wissenschaftliche Güte von IntuiBeat-F bezüglich der Gütekriterien *Gründlichkeit* und *Gültigkeit*, sowie der *zeitlichen Anwendungseffizienz*, als für das Projekt 3D-GUIde wichtiges Kriterium der praktischen Güte, beim Vergleich mit einem als Quasi-Außenkriterium fungierenden Nutzertest mit retrospektivem Think-Aloud-Protokoll überprüft. Als Untersuchungsgegenstand fungierte die Kochwebseite *Chefkoch.de*, bei der es sich auf Basis von Experteneinschätzungen und Nutzereinschätzungen (anhand eines TFQ) um eine eher stärker intuitiv benutzbare Software handelt, bei der es entsprechend zu wenigen, intuitive Benutzung beeinträchtigender Nutzungsproblemen (d.h. reale Nutzungsprobleme) kommen sollte. Wie im Vorgängerexperiment kam derselbe Evaluator und eine zusätzliche Analysesoft-ware während des retrospektiven Interviews zum Einsatz, um dem Evaluator die Zuordnung von Rhythmus-Peaks zu den entsprechenden Videostellen zu erleichtern.

Anhand der statistischen Tests wurde zunächst überprüft, ob sich ein Unterschied bezüglich der Gründlichkeit zu Gunsten von IntuiBeat-F feststellen lässt (Hypothese H1). Anschließend wurde geprüft, ob auch bezüglich Gültigkeit ein Unterschied zu Gunsten von IntuiBeat-F erkennbar ist (Hypothese H2). Abschließend wurde untersucht, inwiefern IntuiBeat-F eine höhere zeitliche Anwendungseffizienz als der Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der formativen Evaluation intuitiver Benutzung der Kochwebseite *Chefkoch.de* aufweisen kann (Hypothese H3). Im folgenden Verlauf werden die Hypothesen bezüglich der Gründlichkeit, der Gültigkeit und der zeitlichen Anwendungseffizienz unter Berücksichtigung der festgestellten Ergebnisse diskutiert, nachdem im Vorfeld auf eine mögliche Konfundierung durch die *Vorerfahrung bei der Nutzung von Kochwebseiten* eingegangen wurde.

6.4.5.1 Überprüfung einer möglichen Konfundierung durch die Vorerfahrung bei der Nutzung von Kochwebseiten

Die Vorerfahrung bei der Nutzung von Kochwebseiten unterschied sich erwartungsgemäß nicht signifikant zwischen den beiden Ausprägungen der unabhängigen Variable *Art der formativen Evaluationsmethode*. Jedoch war die Teststärke aufgrund des geringen beobachteten Effekts und der kleinen Stichprobe gering. Wie bereits zuvor beim fünften Experiment (siehe Abschnitt 6.2), bei dem ebenfalls eine stärker intuitiv benutzbare Webseite formativ bezüglich intuitiver Benutzung evaluiert wurde, kann in diesem Experiment die Operationalisierung der Vorerfahrung durch den TFQ selbst mit hoher Wahrscheinlichkeit in Zusammenhang mit der verwendeten Studierendenstichprobe verantwortlich für dieses Ergebnis sein. Aufgrund der Tatsache, dass eine Kochwebseite ähnlich wie eine Hotelbuchungswebseite, die im Rahmen des fünften Experiments als formativer Untersuchungsgegenstand zum Einsatz kam, von Studierenden in einem eher sporadischen Turnus (d.h. nicht täglich, sondern eher monatlich) genutzt wird, kann der TFQ das in seiner Häufigkeitsdimension nicht sensitiv genug abbilden. Auch die Funktionsdimension lässt ebenfalls nicht viel Varianz zu, da, wie bereits in Teilabschnitt 6.2.5 erwähnt, „alle Features“ (entspricht „4“ auf der Likertskala) das obere Ende der in diesem Zusammenhang verwendeten Likertskala darstellten und sich Studierende mit hoher Wahrscheinlichkeit

auf „genug Features, um damit arbeiten zu können“ (entspricht „2“ auf der Likertskala) einpendeln, wenn sie sich auf die Suche nach Rezepten beschränken.

Dementsprechend ist die praktische Bedeutsamkeit der in diesem Zusammenhang ermittelten Effekte bzw. der Erkenntnisgewinn aufgrund der verwendeten Stichprobe und der mangelnden Sensitivität des TFQ entsprechend eher als unbedeutend einzustufen (siehe Döring & Bortz, 2016), was sich auch in den deskriptiven Unterschieden erkennen lässt. Da darüber hinaus bei der Kontrollvariablen keine signifikanten Korrelationen bezüglich der Gründlichkeit, der Gültigkeit und der zeitlichen Anwendungseffizienz bei beiden formativen Evaluationsmethoden vorlagen ($p > .05$), lag mit hoher Wahrscheinlichkeit keine Konfundierung durch eine unterschiedliche Vorerfahrung bei der Nutzung von Kochwebseiten vor.

6.4.5.2 Überprüfung der Gründlichkeit

Wie erwartet, zeigte sich wie im direkten Vorgängerexperiment im Gegensatz zu den ersten beiden Experimenten eine signifikant höhere Gründlichkeit von IntuiBeat-F gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der Evaluation einer stärker intuitiv benutzbaren Software, sowohl bei Berücksichtigung aller Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F), als auch bei Berücksichtigung nur durch den Algorithmus entdeckter Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F). Der Einsatz einer zusätzlichen Analysesoftware führte mit hoher Wahrscheinlichkeit, wie bereits im Vorgängerexperiment (siehe Teilabschnitt 6.3.5), dazu, dass ein höherer Anteil an Rhythmus-Peaks überhaupt berücksichtigt wurde (d.h. nur rund 30 % der Rhythmus-Peaks waren unberücksichtigt), ein niedriger Anteil von kritischen Ereignissen, die auf reale Nutzungsprobleme hinweisen konnten, unberücksichtigt blieb (d.h. kein kritisches Ereignis blieb bezüglich der realen Nutzungsprobleme unberücksichtigt) und sich in Folge dessen auch die Anzahl an gefundenen realen Nutzungsproblemen durch die strikte Anwendung von IntuiBeat-F (d.h. kein reales Nutzungsproblem blieb unberücksichtigt) nicht veränderte (siehe Teilabschnitt 6.4.4.6). Die von IntuiBeat-F erreichte Gründlichkeit lag bei strikter und weniger strikter Anwendung jedoch nur bei 16 % und die Gründlichkeit des Nutzertests mit retrospektivem Think-Aloud-Protokoll bei 10 % (siehe Teilabschnitt 6.4.4.3). Damit liegen beide Ergebnisse des siebten Experiments eher im unteren Bereich, wenn man sie mit den Ergebnissen aus den Vorgängerexperimenten bzw. speziell aus dem letzten Experiment vergleicht (siehe Teilabschnitt 6.3.5), bei dem ebenfalls eine zusätzliche Analysesoftware zum Einsatz kam.

Beurteilt man diese Ergebnisse des siebten Experiments ebenfalls im Bezug auf die von Koutsabasis et al. (2007) durchführte Meta-Evaluation, mit der bei formativen Evaluationsmethoden für Gebrauchstauglichkeit (d.h. nicht speziell intuitive Benutzung) eine Gründlichkeit im Wertebereich von circa 20 bis 40 % festgestellt werden konnte, liegt die im Rahmen des siebten Experiments beobachtete Gründlichkeit zwar im unteren Wertebereich, das Ergebnis des Nutzertests mit retrospektivem Think-Aloud-Protokoll jedoch noch weitaus niedriger (siehe Teilabschnitt 6.4.4.3). Wie bereits im Rahmen der Diskussion des sechsten Experiments (siehe Teilabschnitt 6.3.5) angesprochen, existieren jedoch unterschiedliche Angaben darüber, wie viele reale Nutzungsprobleme mithilfe von fünf Versuchspersonen entdeckt werden können und welche Gründlichkeit damit realistisch

machbar ist. Nutzt man die im Rahmen des sechsten Experiments vorgestellte Formel (siehe Teilabschnitt 6.3.5), werden bei IntuiBeat-F 11 Versuchspersonen bei einer durchschnittlichen Gründlichkeit von 16 % benötigt, um 85 % aller realen Nutzungsprobleme zu finden. Beim Nutzertest mit retrospektivem Think-Aloud-Protokoll werden bei einer durchschnittlichen Gründlichkeit von 10 %, dafür sogar 18 Versuchspersonen benötigt. Dies liegt zwar außerhalb der von Virzi (1992) und Nielsen und Landauer (1993) vorgeschlagenen „Fünf-Personen-Regel“, aber noch auf einem höheren Niveau als die Arbeit von Spool und Schroeder (2001), in der mit fünf Personen lediglich 35 % aller real vorhandenen Nutzungsprobleme entdeckt werden konnten, da mit IntuiBeat-F hier schon drei Versuchspersonen ausreichen würden und mit dem Nutzertest ebenfalls fünf Versuchspersonen ausreichend sind. Spool und Schroeder (2001) erklären ihre Befunde damit, dass die Gründlichkeit sehr stark von der Gestaltung der Aufgabe abhängt. Besitzt eine Aufgabe sehr viele Lösungsmöglichkeiten, was bei einer Webseite generell der Fall ist, muss jeder Nutzer theoretisch alle verfügbaren Pfade durchlaufen, damit eine hohe Gründlichkeit auch praktisch durch die getestete Methode erreicht werden kann. Da dies sowohl bei CUIs (Experimente 4 und 6) und Webseiten (Experimente 5 und 7) nicht zu erwarten ist, sind die ermittelten Werte bezüglich des Kriteriums der Gründlichkeit als normal einzustufen.

Aufgrund der Tatsache, dass die im Rahmen des sechsten Experiments ermittelte Gründlichkeit von IntuiBeat-F bei der Evaluation einer stärker intuitiv benutzbaren Software zusätzlich signifikant höher und dabei fast doppelt so hoch wie die Gründlichkeit des Nutzertests mit retrospektivem Think-Aloud-Protokoll lag, können zusammenfassend die mit Gründlichkeit in Zusammenhang aufgestellten Hypothesen H1.A und H1.B entsprechend als bestätigt angesehen werden (siehe Tabelle 6.9). In Kombination mit den deskriptiven Ergebnissen des direkten Vorgängerexperiments (siehe Teilabschnitt 6.3.4.6), bei dem ebenfalls die Gründlichkeit von IntuiBeat-F signifikant höher als die Gründlichkeit des Nutzertests mit retrospektivem Think-Aloud-Protokoll lag (siehe Teilabschnitt 6.3.4.3), kann festgehalten werden, dass mithilfe der neuen Analysesoftware mit hoher Wahrscheinlichkeit mehr Rhythmus-Peaks berücksichtigt, weniger kritische Ereignisse, die auf reale Nutzungsprobleme hinweisen konnten, ignoriert und damit letztendlich weniger reale Nutzungsprobleme unberücksichtigt blieben (siehe Teilabschnitt 6.4.4.6). Dies wirkte sich schließlich positiv auf die Gründlichkeit von IntuiBeat-F aus (d.h. es wurde durch die strikte Anwendung von IntuiBeat-F nicht weniger reale Nutzungsprobleme gefunden und so die Gründlichkeit reduziert) und führte dazu, dass die Gründlichkeit bei strikter Anwendung von IntuiBeat-F im Gegensatz zu den ersten beiden Experimenten, sowohl bei der Evaluation von weniger als auch stärker intuitiven Softwareanwendungen, bei IntuiBeat-F stets höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll lag.

6.4.5.3 Überprüfung der Gültigkeit

Wie erwartet, zeigte sich signifikant die höhere Gültigkeit von IntuiBeat-F gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der Evaluation einer stärker intuitiv benutzbaren Software mit zusätzlicher Analysesoftware, sowohl bei Berücksichtigung aller Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F), als auch bei Berücksichtigung nur durch den Algorithmus entdeckter Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F). In beiden Fällen konnte bei IntuiBeat-F eine Gültigkeit

von über 90 % (d.h. weniger strikt: 90 %, strikt: 95 %) und beim Nutzertest mit retrospektivem Think-Aloud-Protokoll eine Gültigkeit von 66 % (d.h. deskriptiver Unterschied von 24 % bzw. 29 %) beobachtet werden (siehe Teilabschnitt 6.4.4.4), was im Vergleich zu anderen formativen Evaluationsmethoden innerhalb der Usability-Forschung (d.h. Evaluationsmethoden, die Gebrauchstauglichkeit formativ messen, aber nicht speziell intuitive Benutzung) ähnlich hoch (d.h. zwischen 74 und 90 % bei empirischen Nutzertests) liegt (siehe Koutsabasis et al., 2007). Im Vergleich zum Vorgängerexperiment, in dem auch die zusätzliche Analysesoftware zum Einsatz kam, lag die Gültigkeit von IntuiBeat-F bei Berücksichtigung aller Nutzungsprobleme (d.h. wenig strikte Anwendung von IntuiBeat-F) mit 90 % leicht höher (d.h. Gültigkeit von 88 % im sechsten Experiment) und ordnet sich im Vergleich zu den ersten beiden Experimenten im oberen Wertebereich ein (d.h. Gültigkeit von 94 % im vierten und Gültigkeit von 79 % im fünften Experiment), weswegen die Gültigkeit in Bezug auf die Arbeit von Koutsabasis et al. (2007) immer noch ein mehr als zufriedenstellendes Ergebnis darstellt.

Betrachtet man die Gültigkeit des siebten Experiments bei ausschließlicher Berücksichtigung von durch Rhythmus-Peaks identifizierten Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F), steigt die Gültigkeit von IntuiBeat-F um 5 % auf 95 %, was darin begründet liegt, dass wie im Vorgängerexperiment, in dem auch eine zusätzliche Analysesoftware zum Einsatz kam, lediglich 3 % aller kritischen Ereignisse unabhängig von Rhythmus-Peaks bei IntuiBeat-F erkannt wurden. Bei den ersten beiden Experimenten, in denen keine solche Software verwendet wurde, wurden noch über 50 % aller kritischen Ereignisse unabhängig von Rhythmus-Peaks protokolliert. Dieser Unterschied kann, wie bereits im Rahmen der Diskussion des Vorgängerexperiments erwähnt (siehe Teilabschnitt 6.3.5), damit erklärt werden, dass der Anteil an realen Nutzungsproblemen an allen Nutzungsproblemen sich dadurch im Vergleich zu den ersten beiden Experimenten weniger stark (d.h. Experiment 4: Steigerung der Gültigkeit um 4 %; Experiment 5: Steigerung der Gültigkeit um 18 %) und damit ähnlich wie im Vorgängerexperiment (d.h. Experiment 6: Steigerung der Gültigkeit um 5 %) veränderte, da fast alle Nutzungsprobleme bis auf eins und alle realen Nutzungsprobleme auch mithilfe von Rhythmus-Peaks entdeckbar waren und damit der Anteil entgangener realer Nutzungsprobleme entsprechend Null war (siehe Teilabschnitt 6.4.4.6). Die 5 % kommen daher nur aufgrund des einen Bewegungsfehlers und individueller Unterschiede zustande, da die Gültigkeit ja auf Versuchspersonenebene berechnet wurde und es hier dennoch zu etwas Varianz gekommen sein kann. Aufgrund der Tatsache, dass im siebten wie im sechsten Experiment eine neue Analysesoftware für die einfachere Zuordnung der Rhythmus-Peaks eingesetzt wurde, sprechen die genannten Zahlen und die geringen Unterschiede bezüglich der Gültigkeit zwischen der Berücksichtigung aller Nutzungsprobleme (d.h. weniger strikte Anwendung von IntuiBeat-F) und der Berücksichtigung ausschließlich mit Rhythmus-Peaks assoziierter Nutzungsprobleme (d.h. strikte Anwendung von IntuiBeat-F) für die Effektivität der neuen Analysesoftware. Nichtsdestotrotz waren die 15 gefundenen Nutzungsprobleme und die 13 realen gefundenen Nutzungsprobleme (siehe Teilabschnitt 6.4.4.6) im Vergleich zu den ersten beiden Experimenten etwas geringer (d.h. Experiment 4: 60 gesamt, 56 real; Experiment 5: 50 gesamt, 47 real), aber im Vergleich zum direkten Vorgängerexperiment, bei dem derselbe Evaluator zum Einsatz kam, auf ähnlichem Niveau (d.h. Experiment 6: 10 gesamt, 9 real).

Wie bereits schon im Rahmen der Diskussion des sechsten Experiments angesprochen (siehe Teilabschnitt 6.3.5), könnte eine mögliche Erklärung für die geringe Anzahl an ent-

deckten (realen) Nutzungsproblemen darin bestehen, dass womöglich im Vergleich zu den vorigen Experimenten aufgrund des im sechsten und siebten Experiment eingesetzten anderen Evaluators weniger Probleme entdeckt werden konnten. Vergleicht man die Anzahl der gefundenen Nutzungsprobleme des sechsten und siebten Experiments mit denen des vierten und fünften Experiments, ist zwar davon auszugehen, dass der Evaluator weniger gefunden hat, jedoch zeigt der Vergleich zwischen IntuiBeat-F und dem Nutzungstest mit retrospektivem Think-Aloud-Protokoll deutlich (siehe Teilabschnitt 6.4.4.4), dass auch in einer solchen Situation der Einsatz von IntuiBeat-F mehr Nutzungsprobleme und sogar mehr reale Nutzungsprobleme bei der Systemnutzung feststellen konnte. Natürlich kann auch der Untersuchungsgegenstand und die damit verbundene Aufgabe dafür verantwortlich sein, dass nicht mehr Nutzungsprobleme identifiziert werden konnten, was man beispielsweise hätte an etwaigen Performanzdaten (z.B. Effektivität bei der Zielerreichung) überprüfen können, die aber aufgrund des qualitativen Fokus des Experiments nicht erhoben wurden. Außerdem kann es bei der Zuordnung der Rhythmus-Peaks zu den entsprechenden Videostellen trotz Analysesoftware zu Fehlern gekommen sein. Ebenso wären Fehler auch bei der softwareunabhängigen Einordnung der Nutzungsprobleme in die handlungsorientierte Fehlertaxonomie von Zapf et al. (1989) möglich, obwohl aufgrund der hohen Beurteilerübereinstimmung eher weniger damit gerechnet werden kann.

Da jedoch die Gültigkeit von IntuiBeat-F beim siebten Experiment stets statistisch signifikant höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll lag (siehe Teilabschnitt 6.4.4.4), kann die Gültigkeit von IntuiBeat-F auch bei Einsatz der zusätzlichen Analysesoftware bei der Evaluation einer stärker intuitiv benutzbaren Software als mehr als zufriedenstellend interpretiert und die in diesem Zusammenhang aufgestellten Hypothesen H2.A und H2.B entsprechend als bestätigt angesehen werden (siehe Tabelle 6.9). In Kombination mit den deskriptiven Ergebnissen des direkten Vorgängerexperiments (siehe Teilabschnitt 6.3.4.6), bei dem ebenfalls die Gültigkeit von IntuiBeat-F signifikant höher als die Gültigkeit des Nutzertests mit retrospektivem Think-Aloud-Protokoll lag, kann festgehalten werden, dass mithilfe der neuen Analysesoftware mit hoher Wahrscheinlichkeit mehr Rhythmus-Peaks berücksichtigt, weniger auf (reale) Nutzungsprobleme hinweisende kritische Ereignisse ignoriert und damit weniger reale Nutzungsprobleme unberücksichtigt blieben. Dies wirkte sich schließlich positiv auf die Gültigkeit aus (d.h. das Verhältnis zwischen den realen und allen Nutzungsproblemen wurde bei Berücksichtigung von ausschließlich mithilfe von Rhythmus-Peaks erkannten kritischen Ereignissen nicht reduziert) und führte dazu, dass diese im Gegensatz zu den ersten beiden Experimenten, sowohl bei der Evaluation von weniger als auch stärker intuitiven Softwareanwendungen, bei IntuiBeat-F stets (d.h. bei strikter und weniger strikter Anwendung) höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll lag.

6.4.5.4 Überprüfung der zeitlichen Anwendungseffizienz

Wie erwartet, zeigte sich die signifikant höhere zeitliche Anwendungseffizienz von IntuiBeat-F gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll auch bei der Evaluation einer stärker intuitiv benutzbaren Software mit Einsatz einer zusätzlichen Analysesoftware signifikant. IntuiBeat-F nahm wie bereits im Vorgängerexperiment (siehe

Teilabschnitt 6.3.5), in dem ebenfalls eine zusätzliche Analysesoftware während des retrospektiven Interviews zum Einsatz kam, sogar circa 40 % weniger Zeit für die Anwendung in Anspruch (siehe Teilabschnitt 6.4.4.5), was aufgrund des sehr geringen Anteils nicht berücksichtigter, auf reale Nutzungsprobleme hinweisender, Rhythmus-Peaks (d.h. lediglich rund 30 %) und der vollständigen Berücksichtigung, auf reale Nutzungsprobleme hinweisender, kritischer Ereignisse auch zu erwarten war (siehe Teilabschnitt 6.4.4.6).

Im Vergleich zum Vorgängerexperiment, bei dem ebenfalls die neue Analysesoftware zum Einsatz kam, konnte mit einer Reduktion von 40 % ein ähnlich beeindruckendes Ergebnis erzielt werden (siehe Teilabschnitt 6.3.4.5). Im Vergleich zu den ersten beiden Experimenten (circa 20 % beim vierten Experiment und circa 25 % beim fünften Experiment gesteigerte zeitliche Anwendungseffizienz im Vergleich zum Nutzertest mit retrospektivem Think-Aloud-Protokoll), bei dem die Zuordnung der Rhythmus-Peaks zu den entsprechenden Videostellen noch vollständig manuell durchgeführt werden musste, konnte durch den Einsatz der neuen Analysesoftware eine Verbesserung der zeitlichen Anwendungseffizienz von rund 15 % erreicht werden, so wie es auch beim direkten Vorgängerexperiment der Fall gewesen ist (siehe Teilabschnitt 6.3.5). Die in diesem Zusammenhang aufgestellte Hypothese H3 kann dementsprechend bei der Evaluation einer stärker intuitiv benutzbaren Software als bestätigt angesehen werden, wenn zusätzlich eine Analysesoftware während des retrospektiven Interviews zum Einsatz kam (siehe Tabelle 6.9). Aufgrund der Tatsache, dass in Kombination mit den deskriptiven Ergebnissen des direkten Vorgängerexperiments (siehe Teilabschnitt 6.3.4.6), in dem ebenfalls die zeitliche Anwendungseffizienz von IntuiBeat-F signifikant höher als die zeitliche Anwendungseffizienz des Nutzertests mit retrospektivem Think-Aloud-Protokoll lag, kann festgehalten werden, dass mithilfe der neuen Analysesoftware mit hoher Wahrscheinlichkeit mehr Rhythmus-Peaks berücksichtigt, weniger, auf reale Nutzungsprobleme hinweisende, kritische Ereignisse ignoriert und damit weniger reale Nutzungsprobleme unberücksichtigt blieben (siehe Teilabschnitt 6.4.4.6). Dies wirkte sich schließlich positiv auf die zeitliche Anwendungseffizienz aus und führte dazu, dass diese im Gegensatz zu den ersten beiden Experimenten, sowohl bei der Evaluation von weniger als auch stärker intuitiv benutzbaren Softwareanwendungen, bei IntuiBeat-F stets (d.h. bei strikter und weniger strikter Anwendung) höher als beim Nutzertest mit retrospektivem Think-Aloud-Protokoll lag.

6.4.6 Schlussfolgerung

Zusammenfassend kann festgehalten werden, dass die wissenschaftliche Güte von IntuiBeat-F als formative Evaluationsmethode für intuitive Benutzung hinsichtlich des Gütekriteriums *Gültigkeit* auch bei der Untersuchung einer stärker intuitiv gestalteten Software bei Verwendung einer zusätzlichen Analysesoftware während des retrospektiven Interviews empirisch, sowohl bei strikter als auch weniger strikter Anwendung, bestätigt werden konnte (siehe Tabelle 6.9). Darüber hinaus konnte IntuiBeat-F auch ein für das Projekt 3D-GUIde wichtiger Teilaspekt praktischer Güte aufgrund seiner im Vergleich zum Nutzertest mit retrospektivem Think-Aloud-Protokoll höheren zeitlichen Anwendungseffizienz empirisch zugesprochen werden (siehe Tabelle 6.9). Konfundierungen durch Unterschiede in der Vorerfahrung bei der Nutzung von CUIs können mit hoher Wahrscheinlichkeit ausgeschlossen werden. Die Effekt- und Teststärken lagen dabei überwiegend im oberen Bereich.

Tabelle 6.9. *Übersicht der mithilfe des siebten Experiments bestätigten Hypothesen im Zuge der Meta-Evaluation von IntuiBeat-F (KEs: kritische Ereignisse; RP-KEs: mit Rhythmus-Peaks assoziierte kritische Ereignisse).*

Hypothese	Experiment 6
(H1) Überprüfung der Gründlichkeit:	
- (A) Alle KEs	✓
- (B) Nur RP-KEs	✓
(H2) Überprüfung der Gültigkeit:	
- (A) Alle KEs	✓
- (B) Nur RP-KEs	✓
(H3) Überprüfung der zeitlichen Anwendungseffizienz	✓

Wie bereits im direkten Vorgängerexperiment und im Gegensatz zu den beiden ersten Experimenten, in denen keine zusätzliche Analysesoftware zum Einsatz kam und der Evaluator mit hoher Wahrscheinlichkeit, aufgrund der komplizierten Zuordnung von Rhythmus-Peaks zu kritischen Ereignissen, Rhythmus-Peaks nicht berücksichtigt hat, konnte auch die wissenschaftliche Güte von IntuiBeat-F bezüglich des Gütekriteriums der Gründlichkeit sowohl bei strikter als auch weniger strikter Anwendung bestätigt werden (siehe Tabelle 6.9). Zusammenfassend konnte dementsprechend durch den Einsatz der neuen Analysesoftware bei den letzten beiden Experimenten, in denen eine weniger und eine stärker intuitive Software evaluiert wurde, die wissenschaftliche und die praktische Güte von IntuiBeat-F als formative Evaluationsmethode für intuitive Benutzung nachgewiesen werden.

6.5 Allgemeine Diskussion und Zusammenfassung

Das Ziel des vorliegenden Kapitels war es erstmalig mithilfe von vier Experimenten zu demonstrieren, dass die Evaluationsmethode *IntuiBeat-F*, die eine Adaption der von Park und Brünken (2015) entwickelten Rhythmusmethode für die formative Evaluation intuitiver Benutzung im HCI-Bereich darstellt, formative Evaluationsergebnisse mit wissenschaftlicher Güte auf zeitlich effiziente Weise bereitstellen kann (siehe Tabelle 6.10). Wie in Abschnitt 4.2 angesprochen, bietet die Rhythmusmethode (siehe Korbach et al., 2018; Park & Brünken, 2015), aufgrund ihrer Möglichkeit mentale Beanspruchung objektiv anhand von Inhibitionsprozessen erfassen zu können, nicht nur das Potential intuitive Benutzung summativ evaluieren zu können (d.h. wissenschaftliche und praktische Güte wurde bereits im letzten Kapitel demonstriert), sondern kann auch als rhythmische Inhibitionsaufgabe mit individueller Baseline für die formative Evaluation intuitiver Benutzung genutzt werden. Mithilfe des AMPD-Algorithmus (siehe Scholkmann et al., 2012) kann der Evaluator bei der Identifikation intuitive Benutzung beeinträchtigender Nutzungsprobleme (d.h. reale Nutzungsprobleme, die auf Basis der handlungsorientierten Fehlertaxonomie bestimmt wurden, siehe Abschnitt 3.2.3) mit hoher Wahrscheinlichkeit von seiner eigenen Willkürlichkeit entkoppelt werden (d.h. Evaluator widmet sich aufgrund seines Expertenwissens unnötigerweise nicht realen Nutzungsproblemen), indem ihm der Algorithmus lediglich anhand sog. Rhythmus-Peaks „sichere“ Hinweise auf potentielle reale Nutzungsprobleme (d.h. kritische Ereignisse) bereitstellt.

Tabelle 6.10. *Übersicht der mithilfe der Experimente 4 bis 7 bestätigten Hypothesen im Zuge der Meta-Evaluation von IntuiBeat-F (KEs: kritische Ereignisse; RP-KEs: mit Rhythmus-Peaks assoziierte kritische Ereignisse).*

Hypothese	Experiment 4	Experiment 5	Experiment 6	Experiment 7
(H1) Überprüfung der Gründlichkeit:				
- (A) Alle KEs	✓	✓	✓	✓
- (B) Nur RP-KEs	✗	✗	✓	✓
(H2) Überprüfung der Gültigkeit:				
- (A) Alle KEs	✓	✓	✓	✓
- (B) Nur RP-KEs	✓	✓	✓	✓
(H3) Überprüfung der zeitlichen Anwendungseffizienz	✓	✓	✓	✓

Da mittlere Rhythmusabweichungen im summativen Bereich als potentielle Indikatoren für intuitive Benutzung bereits qualifiziert waren (siehe Kapitel 5), ist anzunehmen, dass die mithilfe des Algorithmus identifizierten echten lokalen Extrema mit hoher Wahrscheinlichkeit auf kritische Ereignisse hinweisen, die aufgrund hoher mentaler Beanspruchung und damit durch ein Defizit in der intuitiven Benutzung mit dem System verursacht wurden. Dementsprechend sollte IntuiBeat-F im Vergleich mit dem in Abschnitt 3.5 als geeignetes objektives Quasi-Außenkriterium identifizierten Nutzertest mit retrospektivem Think-Aloud-Protokoll eine höhere wissenschaftliche Güte und zeitliche Anwendungseffizienz aufweisen, wenn es sich um eine „bessere“ formative Evaluationsmethode für intuitive Benutzung handelt. Die Ergebnisse der vier Experimente, die in diesem Kapitel beschrieben wurden, konnten erwartungsgemäß bestätigen, dass IntuiBeat-F wissenschaftliche und zeitliche Anwendungseffizienz, als für das Anwenderprojekt 3D-GUIde wichtiger Aspekt praktischer Güte, als objektive Methode zur formativen Evaluation bezüglich des Ausmaßes an intuitiver Benutzung, sowohl bei weniger als auch stärker intuitiv benutzbaren Softwareanwendungen, fungieren kann. Beide in diesem Zusammenhang aufgestellten Forschungsfragen (Forschungsfragen 2 und 3) konnten demzufolge beantwortet werden. Im Folgenden werden die einzelnen Befunde bezüglich des Nachweises wissenschaftlicher Güte und zeitlicher Anwendungseffizienz zusammengefasst betrachtet und kritisch diskutiert, einhergehend mit Bezugnahme auf die Forschungsliteratur.

Einschlägige Forschungsliteratur im Bereich der Meta-Evaluation von formativen Evaluationsmethoden (z.B. Hartson et al., 2001; Koutsabasis et al., 2007; Sears, 1997) empfiehlt zunächst die wissenschaftliche Güte einer Methode hinsichtlich Gründlichkeit und Gültigkeit nachzuweisen, bevor sich der Zuverlässigkeit gewidmet werden soll. Hartson et al. (2001) argumentieren beispielsweise, dass sich der Entwickler einer Methode erstmal darüber klar werden muss, wie seine Methode üblicherweise angewendet werden soll. Soll eine Methode, so wie es bei empirischen Methoden (z.B. Nutzertest) im Forschungsgebiet der Gebrauchstauglichkeit Usus ist, von einer Gruppe von Evaluatoren angewendet werden (d.h. gefundene Nutzungsprobleme werden üblicherweise über mehrere Evaluatoren hinweg aggregiert, siehe Hartson et al., 2001), ist diese Tatsache unbedingt beim Priorisieren der Zuverlässigkeit als wissenschaftliches Gütekriterium zu berücksichtigen.

Eine hohe Zuverlässigkeit einer formativen Evaluationsmethode wird normalerweise durch Standardisierung erreicht und schränkt dadurch die Sichtweise des Evaluators bzw. die individuelle Variabilität in der Entdeckung von Nutzungsproblemen künstlich ein. Die Gründlichkeit, die von mehreren Evaluatoren erhoben wird, wird jedoch dadurch reduziert, dass sich diese aufgrund der starken Standardisierung mit höherer Wahrscheinlichkeit den gleichen Bedienelementen und damit verbundenen Nutzungsproblemen widmen. Soll hingegen eine Methode vorwiegend individuell eingesetzt werden (z.B. bei einigen Inspektionsmethoden wie der heuristischen Evaluation denkbar, obwohl diese auch üblicherweise in Gruppen angewendet und Nutzungsproblemen über mehrere Evaluatoren hinweg aggregiert werden, siehe Hartson et al., 2001), ist eine individuelle Steigerung der Zuverlässigkeit auf Kosten der Gründlichkeit vielleicht wünschenswert. Hartson et al. (2001) empfehlen deswegen immer erst Gründlichkeit und Gültigkeit zu priorisieren, da zumindest die Zuverlässigkeit über eine Gruppe von Evaluatoren sich dadurch oft automatisch verbessert. Es wurde sich dementsprechend im Rahmen dieser Arbeit entschieden lediglich Gründlichkeit und Gültigkeit als Indikatoren wissenschaftlicher Güte empirisch zu untersuchen.

Diese Entscheidung stellt zwar eine Limitation der vorliegenden Arbeit dar und es sollte sich in künftigen Studien der Untersuchung der Zuverlässigkeit gewidmet werden, dennoch ist bei der Anwendung von IntuiBeat-F zumindest theoretisch von einer hohen individuellen Zuverlässigkeit auszugehen. Dafür spricht, dass, aufgrund der durch den AMPD-Algorithmus vorgeschlagenen Rhythmus-Peaks, die Evaluatoren standardisiert, auf reale Nutzungsprobleme indizierende, kritische Ereignisse hingewiesen werden und man über alle Experimente hinweg eine ähnliche Gründlichkeit bei der ausschließlichen Berücksichtigung von Rhythmus-Peaks (d.h. strikte Anwendung von IntuiBeat-F) bei der Ableitung von Nutzungsproblemen durch ähnlich qualifizierte voneinander verschiedene Evaluatoren (d.h. Evaluator des vierten und fünften Experiments war ein Bachelorstudent mit hoher Kenntnis des Untersuchungsgegenstands, aber wenig Erfahrung bei der Durchführung; Evaluator des sechsten und siebten Experiments war eine Bachelorstudentin mit hoher Kenntnis des Untersuchungsgegenstands, aber wenig Erfahrung bei der Durchführung) erkennen konnte. Vergleicht man zusätzlich die ermittelten Werte der letzten beiden Experimente mit den Werten der ersten beiden Experimente bei Berücksichtigung aller Nutzungsprobleme bezüglich der Gründlichkeit (d.h. weniger strikte Anwendung von IntuiBeat-F), stellt man fest, dass womöglich durch die neue Analysesoftware (Experiment 6 und 7) eine höhere Standardisierung bzw. Individualzuverlässigkeit erreicht wurde, die zu Lasten der Gründlichkeit des Evaluators ging und sich mit hoher Wahrscheinlichkeit auch in einer Gruppenevaluation (d.h. Gruppen-Zuverlässigkeit) bemerkbar machen sollte (siehe Tabelle 6.11). Zukünftige Studien sollten sich demzufolge dieser Problematik widmen und gezielt überprüfen, wie sich die Gruppenzuverlässigkeit und Individualzuverlässigkeit bei IntuiBeat-F gestaltet, wenn verschiedene Evaluatoren mit unterschiedlicher Erfahrung die formative Evaluation intuitiver Benutzung mithilfe von IntuiBeat-F durchführen.

Das wissenschaftliche Gütekriterium der Gründlichkeit von IntuiBeat-F konnte in allen vier Experimenten bei Berücksichtigung aller, auf reale Nutzungsprobleme hinweisenden, kritischen Ereignisse nachgewiesen (siehe Tabelle 6.10) und die damit verbundene Hypothese H1.A bei einer weniger strikten Anwendung von IntuiBeat-F bestätigt werden ($d \geq .8$). Bei Berücksichtigung ausschließlich mit Rhythmus-Peaks assoziierter, auf reale Nutzungsprobleme hinweisender, kritischer Ereignisse und damit der strikten Anwendung von

IntuiBeat-F konnte jedoch nur bei den letzten beiden Experimenten, in denen während des retrospektiven Interviews eine zusätzliche Analysesoftware zur Unterstützung des Evaluators für die Zuordnung der Rhythmus-Peaks zu den entsprechenden Videostellen zum Einsatz kam, ein signifikanter Unterschied zwischen IntuiBeat-F und dem Nutzertest mit retrospektivem Think-Aloud-Protokoll festgestellt werden. Eine Übersicht der deskriptiven Ergebnisse ist Tabelle 6.11 zu entnehmen.

Aufgrund der Tatsache, dass im vierten und fünften Experiment der Anteil nicht berücksichtigter Rhythmus-Peaks, der Anteil nicht berücksichtigter kritischer Ereignisse und der Anteil entgangener Nutzungsprobleme, sowohl bezogen auf alle mit Rhythmus-Peaks identifizierten Nutzungsprobleme als auch beschränkt auf reale Nutzungsprobleme, verringert wurde (siehe Tabelle 6.12) und sich gleichzeitig deskriptiv ein geringerer Unterschied bezüglich der Trefferrate bzw. Fehlalarmrate zwischen der Berücksichtigung aller kritischen, auf reale Nutzungsprobleme hinweisenden, Ereignisse und der Berücksichtigung nur mit Rhythmus-Peaks assoziierten, auf reale Nutzungsprobleme hinweisenden, kritischer Ereignisse zeigte (siehe „Trefferrate reale NPs“ und „Fehlalarmrate reale NPs“ in Tabelle 6.11), ist mit hoher Wahrscheinlichkeit davon auszugehen, dass der Evaluator unabhängig von den generierten Rhythmus-Peaks kritische Ereignisse notierte, woraus sich reale Nutzungsprobleme ableiteten. Diese Nutzungsprobleme gingen jedoch nicht in die Berechnung der Gründlichkeit ein, wenn ausschließlich, auf reale Nutzungsprobleme hinweisende, kritische mit Rhythmus-Peaks verknüpfte Ereignisse berücksichtigt wurden. Der Evaluator nutzte dementsprechend bei der weniger strikten Anwendung von IntuiBeat-F wirklich kaum die Möglichkeiten die IntuiBeat-F bei einer strikten Anwendung zu bieten hat.

Demzufolge ist die Gründlichkeit von IntuiBeat-F bei ausschließlicher Berücksichtigung von Rhythmus-Peaks geringer und die des Nutzertests etwas höher, da dieser weiterhin die gleiche Anzahl an realen Nutzungsproblemen entdeckt und die Gründlichkeit ja immer im Verhältnis zu allen, während der Interaktion auftretenden, realen Nutzungsproblemen berechnet wird, welche ja die Vereinigung der realen Nutzungsprobleme von IntuiBeat-F und des Nutzertests mit retrospektivem Think-Aloud-Protokoll darstellt (siehe Teilabschnitt 3.2.3). Diese Begründung für die verringerte Gründlichkeit im vierten und fünften Experiment erscheint noch wahrscheinlicher, wenn man sich die Ergebnisse der beiden Experimente im Vergleich zum sechsten und siebten Experiment anschaut, bei denen es bezüglich der Trefferrate bzw. Fehlalarmrate deskriptiv viel geringere Unterschiede zwischen der Berücksichtigung aller kritischen Ereignisse (d.h. weniger strikte Anwendung) und ausschließlich mithilfe von Rhythmus-Peaks identifizierter kritischer Ereignisse (d.h. strikte Anwendung) gab. Der Evaluator protokollierte hier mit hoher Wahrscheinlichkeit aufgrund der zusätzlichen Unterstützung durch die neue Analysesoftware keine, auf reale Nutzungsprobleme hinweisenden, kritischen Ereignisse unabhängig von Rhythmus-Peaks, was die Daten auch bestätigen konnten (siehe Tabelle 6.12). Der Einsatz einer neuen Analysesoftware konnte eine striktere Anwendung von IntuiBeat-F fördern und den Evaluator auf diese Weise dabei unterstützen, durch IntuiBeat-F im Vergleich zum Nutzertest mit retrospektivem Think-Aloud-Protokoll eine signifikant höhere Gründlichkeit bei der Evaluation einer weniger intuitiv benutzbaren Software (CUI: Experiment 6) und einer stärker intuitiv benutzbaren Software (Webseite: Experiment 7) zu erzielen (siehe „Gründlichkeit“ in Tabelle 6.11). Die damit verbundene Hypothese H1.B konnte mithilfe der letzten beiden Experimente also auch bestätigt werden ($d \geq .8$).

Tabelle 6.11. *Übersicht allgemeiner deskriptiver Daten der Experimente 4 bis 7. I = IntuiBeat-F; N = Nutzertest mit retrospektivem Think-Aloud-Protokoll (NPs: Nutzungsprobleme; KEs: kritische Ereignisse; RP-KEs: kit Rhythmus-Peaks assoziierte kritische Ereignisse).*

Metrik	Experiment 4		Experiment 5		Experiment 6		Experiment 7	
	I	N	I	N	I	N	I	N
Trefferrate NPs (%):								
- Alle NPs aus KEs	93.09	88.26	79.91	76.56	93.94	88.24	94.12	92.31
- Nur NPs aus RP-KEs	100	88.26	100	76.56	100	88.24	100	92.31
Trefferrate reale NPs (%):								
- Alle NPs aus KEs	88.16	87.05	77.68	75.52	84.85	26.47	82.35	57.69
- Nur NPs aus RP-KEs	95.75	87.05	97.40	75.52	93.33	26.47	90.32	57.69
Fehlalarmrate NPs (%):								
- Alle NPs aus KEs	6.91	11.74	20.09	23.44	6.06	11.74	5.88	7.69
- Nur NPs aus RP-KEs	0	11.74	0	23.44	0	11.74	0	7.69
Fehlalarmrate reale NPs (%):								
- Alle NPs aus KEs	11.84	12.95	22.32	24.48	15.15	73.53	17.65	42.31
- Nur NPs aus RP-KEs	4.25	12.95	2.60	24.48	6.67	73.53	9.68	42.31
Anzahl NPs:								
- Alle NPs aus KEs	60	54	50	37	10	6	15	8
- Nur NPs aus RP-KEs	41	54	37	37	10	6	14	8
Anzahl realer NPs:								
- Alle NPs aus KEs	56	51	47	36	9	1	13	3
- Nur NPs aus RP-KEs	40	51	36	36	9	1	13	3
Anteil realer NPs an NPs (%):								
- Alle NPs aus KEs	93.33	94.44	94	97.30	90	16.67	86.67	37.50
- Nur NPs aus RP-KEs	97.56	94.44	97.30	97.30	90	16.67	92.86	37.50
Gründlichkeit (%):								
- Alle NPs aus KEs	32	22	29	19	27	18	16	10
- Nur NPs aus RP-KEs	19	23	14	20	27	18	16	10
Gültigkeit (%):								
- Alle NPs aus KEs	92	81	79	64	88	51	90	66
- Nur NPs aus RP-KEs	96	81	97	64	90	51	95	66
Zeitliche Anwendungseffizienz (s)	193.84	247.51	268.59	358.77	421.96	721.33	352.50	603.89

Das wissenschaftliche Gütekriterium der Gültigkeit von Intuibeat-F konnte in allen Experimenten sowohl bei Berücksichtigung aller, auf Nutzungsprobleme hinweisenden, kritischen Ereignisse (d.h. weniger strikte Anwendung von IntuiBeat-F) als auch ausschließlich mit Rhythmus-Peaks assoziierten kritischen Ereignisse (d.h. strikte Anwendung von IntuiBeat-F) nachgewiesen werden ($d \geq .8$). Die beiden damit verbundenen Hypothesen H2.A und H2.B konnten somit sowohl bei der Evaluation einer wenig intuitiv benutzbaren Software (Experiment 4 und 6) und einer stärker intuitiv benutzbaren Software (Experiment 5 und 7) bestätigt werden (siehe Tabelle 6.10). Vergleicht man die Ergebnisse bei Berücksichtigung aller kritischen Ereignisse und ausschließlich der mit Rhythmus-Peaks assoziierten kritischen Ereignisse miteinander (siehe „Gültigkeit“ in Tabelle 6.11), stellt man fest, dass man in allen vier Experimenten eine höhere Gültigkeit erreicht, wenn sich der Evaluator auf die Rhythmus-Peaks verlässt und damit IntuiBeat-F strikt anwendet, da Rhythmus-Peaks mit hoher Wahrscheinlichkeit auf reale Nutzungsprobleme hinweisen.

Bei der vollständig manuellen Zuordnung der Rhythmus-Peaks zu den Videostellen (Experiment 4 und 5) ließ sich aufgrund des höheren Anteils nicht berücksichtigter Rhythmus-

Peaks, des höheren Anteils nicht berücksichtigter kritischer Ereignisse und des höheren Anteils entgangener Nutzungsprobleme, sowohl bezogen auf alle mit Rhythmus-Peaks identifizierten Nutzungsprobleme als auch beschränkt auf reale Nutzungsprobleme (siehe Tabelle 6.12), eine höhere Steigerung der Gültigkeit durch die strikte Anwendung von IntuiBeat-F nicht erkennen (vor allem im fünften Experiment). Mithilfe der neuen Analysesoftware konnte dieser Unterschied etwas konstanter gehalten und damit eine stärker valide Zuordnung von Rhythmus-Peaks zu, auf reale Nutzungsprobleme hinweisenden, Ereignissen unterstützt werden (siehe „Trefferrate reale NPs“ und „Fehlalarmrate reale NPs“ in Tabelle 6.11). Nichtsdestotrotz ist dieser Einfluss nicht so präsent wie bei der Gründlichkeit (d.h. wird insbesondere beim Vergleich des vierten und siebten Experiments deutlich), da IntuiBeat-F erwartungsgemäß die Trefferrate bzw. Fehlalarmrate stärker bezüglich realer Nutzungsprobleme steigert bzw. reduziert. Damit ist IntuiBeat-F auch schon bei manueller Zuordnung der Rhythmus-Peaks und damit unabhängig von der Analysesoftware fähig, eine signifikant höhere Gültigkeit im Vergleich zum Nutzertest mit retrospektivem Think-Aloud-Protokoll zu erreichen.

Neben der wissenschaftlichen Güte konnte IntuiBeat-F auch ein, für das Projekt 3D-GUIde wichtiger Aspekt praktischer Güte, aufgrund einer höheren zeitlichen Anwendungseffizienz gegenüber dem Nutzertest mit retrospektivem Think-Aloud-Protokoll in allen vier Experimenten nachgewiesen werden ($d \geq .8$). Die in diesem Zusammenhang aufgestellte Hypothese H3 konnte damit sowohl bei der Evaluation einer wenig intuitiv benutzbaren Software (Experiment 4 und 6) und einer stärker intuitiv benutzbaren Software (Experiment 5 und 7) bestätigt werden (siehe Tabelle 6.10). Obwohl durch die Einführung der neuen Analysesoftware, eine höhere reale Trefferrate bzw. niedrigere Fehlalarmrate (siehe Tabelle 6.11), ein geringerer Anteil nicht berücksichtigter kritischer Ereignisse, ein höherer Anteil berücksichtigter Rhythmus-Peaks (siehe Tabelle 6.12) und damit ein größerer signifikanter Unterschied zwischen IntuiBeat-F und dem Nutzertest mit retrospektivem Think-Aloud-Protokoll beobachtbar ist, ist anzumerken, dass bei den letzten beiden Experimenten nur sehr wenige reale Nutzungsprobleme im Vergleich zu den ersten beiden Experimenten (d.h. deskriptiv weniger als ein Drittel) festgestellt werden konnten, weswegen die Zeit für die Identifikation eines realen Nutzungsproblems vergleichsweise so hoch ausfiel (siehe „Zeitliche Anwendungseffizienz“ in Tabelle 6.11).

Obwohl, auf Basis der Ergebnisse der durchgeführten vier Experimente, IntuiBeat-F unter Verwendung einer Analysesoftware während des retrospektiven Interviews wissenschaftliche und zeitliche Anwendungseffizienz attestiert werden kann, sollten bei der Interpretation dieser Ergebnisse eine Reihe methodischer Limitierungen und Alternativerklärungen berücksichtigt werden.

Zunächst wurden als Untersuchungsgegenstände lediglich zwei Arten von Softwareanwendungen (d.h. CUIs und Webseiten) und eine Auswahl an damit verbundenen Aufgaben verwendet, die auf Basis von Expertenschätzungen ausgewählt wurden. Es wurde sich für diese Art von Softwareanwendungen entschieden, da sie beide dadurch gekennzeichnet sind, nicht vollständig lineare Interaktionsabläufe zu bieten, sondern aufgrund der Vielzahl an möglichen Lösungswegen und der damit verbundenen hohen Evaluationskomplexität gute Untersuchungsgegenstände für eine Meta-Evaluation darstellen (CUI: siehe Akers et al. (2012); Webseiten: siehe Barnum (2010)). Jedoch erfolgte die Entscheidung für CUIs und die damit verbundene Auswahl konkreter CUIs bzw. Aufgaben nicht vollständig frei

und war maßgeblich durch das Projekt 3D-GUIde motiviert, um auf Basis der Ergebnisse Interaktionspattern entwickeln zu können. Dabei stellten die CUIs im Vergleich zu den Webseiten immer die weniger intuitiven Softwareanwendungen dar, aus denen möglichst viele reale Nutzungsprobleme abgeleitet werden sollten.

Tabelle 6.12. *Übersicht über durch ausschließliche Berücksichtigung von Rhythmus-Peaks (d.h. strikte Anwendung von IntuiBeat-F) entgangene kritische Ereignisse und Nutzungsprobleme in den Experimenten 4 bis 7 (NPs: Nutzungsprobleme; KEs: kritische Ereignisse; RP-KEs: mit Rhythmus-Peaks assoziierte kritische Ereignisse). Alle Angaben erfolgten in Prozent.*

Metrik	Experiment 4	Experiment 5	Experiment 6	Experiment 7
Anteil nicht berücksichtigter RPs (Alle NPs)	41.25	31.86	23.08	32.61
Anteil nicht berücksichtigter RPs (Reale NPs)	43.75	33.63	28.20	39.13
Anteil nicht berücksichtigter KEs (Alle NPs)	66.78	56.98	3.23	3.13
Anteil nicht berücksichtigter KEs (Reale NPs)	66.42	56.90	0	0
Anteil entgangener NPs	31.67	26	0	6.67
Anteil entgangener realer NPs	28.57	23.40	0	0

Durch diese Einschränkungen ist natürlich eine Generalisierung der Ergebnisse für weitere Untersuchungsgegenstände schwierig, wobei dies auch gar nicht der Fokus der vorliegenden Arbeit war. Bevor Ergebnisse auf anderen Domänen, Aufgaben und Populationen übertragen werden können, sollte die wissenschaftliche Güte eines Verfahrens zunächst für eine Domäne nachgewiesen werden (siehe Sedlmeier & Renkewitz, 2008). Durch die Beschränkung auf wenige Domänen und Aufgaben konnte man zunächst die wissenschaftliche Güte von IntuiBeat-F und die dahinter stehende Theorie erstmalig im Bereich intuitiver Benutzung verifizieren, was ein übliches Vorgehen im Bereich formativer Meta-Evaluation (siehe Hartson et al., 2001) und bei einer Meta-Evaluation im Allgemeinen, sowohl im Feld (siehe Flick, 2010; Hyman, 1982) als auch im Labor (siehe Bracht & Glass, 1968; Sedlmeier & Renkewitz, 2008) darstellt, bevor Ergebnisse generalisiert werden sollten. Das Projekt 3D-GUIde bot durch seine Beschränkungen damit den idealen Rahmen für eine erste Meta-Evaluation von IntuiBeat-F. Da das Ziel dieser Arbeit vielmehr darin bestand eine formative Evaluationsmethode für intuitive Benutzung zu entwickeln, die zeitlich effizient bei User Interfaces (speziell 3D-CUIs) eingesetzt werden kann, um in kurzer Zeit möglichst viele potentielle Informationen für die Entwicklung 3D-CUI-Interaktionspatterns zu sammeln, liefern die in diesem Kapitel vorgestellten Ergebnisse wichtige erste Erkenntnisse bezüglich der Güte von IntuiBeat-F und das im Rahmen dieser Arbeit gesteckte Ziel kann als erfüllt betrachtet werden.

Zusätzlich wurde sich bei der Meta-Evaluation dafür entschieden, die Anzahl aller verfügbaren (realen) Nutzungsprobleme durch die Kardinalität der Vereinigung der gefundenen (realen) Nutzungsprobleme von beiden getesteten formativen Evaluationsmethoden zu approximieren. Dieses Vorgehen stellt laut Hartson et al. (2001), Koutsabasis et al. (2007) und Sears (1997) eine gute Möglichkeit dar, da neben der eigentlichen Evaluation des Untersuchungsgegenstands mithilfe der beiden formativen Evaluationsmethoden kein zu-

sätzlicher Aufwand nötig ist. Diesen Vorteil erkauft man sich jedoch dadurch, dass dies lediglich eine Annäherung an den echten Totalwert ist und man durch dieses Vorgehen bei der Berechnung der Gültigkeit Folgendes beachten muss. Die realen Nutzungsprobleme sollten hierbei nicht ins Verhältnis mit der Vereinigung gesetzt werden, da diese ja auch Probleme enthält, die von der gerade getesteten Methode gar nicht gefunden wurden. Da es bei der Gültigkeit aber darum geht, herauszufinden, welcher Anteil der mit einer bestimmten Methode gefundenen Nutzungsprobleme überhaupt real ist, sollten stattdessen nur die Nutzungsprobleme der gerade getesteten Methode als Referenzmenge genutzt werden (siehe Koutsabasis et al., 2007).

Wie bereits erwähnt, stellt dieses Vorgehen zudem nur eine Abschätzung und damit nicht den Königsweg dar. Laut Hartson et al. (2001) bietet das sog. „Salting“ bzw. „Seeding“ des Untersuchungsgegenstands die größte Kontrolle über die entdeckbaren Nutzungsprobleme. Hierbei wird vom Evaluator ein System entwickelt oder ein System als Untersuchungsgegenstand ausgewählt, in dem sich keine (realen) Nutzungsprobleme befinden. Der Evaluator fügt dann selbst (reale) Nutzungsprobleme ein und weiß somit die exakte Anzahl und die Umstände des Auftretens jedes eingefügten (realen) Nutzungsproblems. Jedoch kann dies laut Hartson et al. (2001) auch problematisch sein, da erstmal ein entsprechendes System bereitgestellt werden muss und das Hinzufügen von Nutzungsproblemen sehr stark vom Vorwissen des Evaluators abhängt, was auf Kosten der externen bzw. ökologischen Gültigkeit geht und damit die Übertragbarkeit auf das Alltagsgeschäft erschwert. Laut Hartson et al. (2001) wissen Usability-Experten genau, dass die durchs „Salting“ erhaltenen Ergebnisse nicht die Variabilität, die Überraschungen und tatsächliche Echtheit von Nutzungsproblemen in einem Nutzertest in der Praxis widerspiegeln können.

Da „Salting“ im Rahmen des Projekts 3D-GUIde nur schwer machbar gewesen wäre, da Interaktionspatterns ja basierend auf echten Nutzungsproblemen aus dem Alltag der Nutzer entwickelt werden sollten und man aufgrund der Beschränkung auf zwei Arten von Untersuchungsgegenständen (d.h. CUIs und Webseiten) eh eine geringe externe Gültigkeit besaß, wurde sich gegen eine weitere Reduzierung der Generalisierbarkeit durch „Salting“ und damit für die Vereinigung der (realen) Nutzungsprobleme beider getesteter Methoden als Referenzmenge entschieden und auf diese Weise ein möglichst praxisnaher Vergleich im Bereich der CUIs und Webseiten ermöglicht.

Eine weitere Limitation der vorliegenden Meta-Evaluation von IntuiBeat-F kann auch die Tatsache sein, dass bei jedem Experiment beide Evaluationsmethoden jeweils vom selben Evaluator durchgeführt wurden, wobei ein Evaluator die Experimente 4 und 5 durchführte und ein anderer Evaluator die Experimente 6 und 7. Auf diese Weise kann es zu Übertragungseffekten beim Evaluator trotz Randomisierung gekommen sein, da ihm aus einer Evaluationsmethode bekannte Nutzungsprobleme mit hoher Wahrscheinlichkeit auch bei der anderen Evaluationsmethode leichter aufgefallen sind. Ein Indiz dafür kann sein, dass über alle vier Experimente hinweg ein großer Anteil von Nutzungsproblemen gefunden wurde, der in beiden getesteten Methoden identifiziert werden konnte. Nichtsdestotrotz wurden in beiden Methoden auch einzigartige Nutzungsprobleme entdeckt, die nicht in der anderen Methode entdeckt wurden. Insbesondere die höhere Anzahl entdeckter realer Nutzungsprobleme bei IntuiBeat-F führte zu dessen höherer Gültigkeit und Gründlichkeit im Vergleich zum Nutzertest mit retrospektivem Think-Aloud-Protokoll.

Wie bereits in der Einführung dieses Kapitels angesprochen, stellt der sog. Evaluatoreffekt, also die Tatsache, dass verschiedene Evaluatoren zu unterschiedlichen Nutzungsproblemen bei der Untersuchung desselben Systems kommen können (Hertzum & Jacobsen, 2001) im Bereich der Usability-Forschung ein zusätzliches Problem dar (z.B. Hertzum & Jacobsen, 2001; Hertzum et al., 2014; Hornbæk & Frøkjær, 2008; Jacobsen et al., 1998). Laut einer Studie von Hornbæk und Frøkjær (2008), in der 50 Evaluatoren (d.h. wenig Vorerfahrung bei der Durchführung von Nutzertests so wie in den vier hier vorgestellten Experimenten) und ihre entdeckten Nutzungsprobleme miteinander verglichen wurden, konnte lediglich eine Übereinstimmung von 40 % beobachtet werden, was für einen starken Effekt des Evaluators spricht, da unter eigentlich gleichen Ausgangsbedingungen unterschiedliche Nutzungsprobleme abgeleitet wurden. Deswegen wurde zur Kontrolle des Effekts der gleiche Evaluator in beiden Bedingungen verwendet und beide Bedingungen über alle Versuchspersonen hinweg randomisiert, um Übungseffekte und den Evaluatoreffekt in beiden Gruppen konstant zu halten. Mithilfe dieses Vorgehens konnte überhaupt zum ersten Mal ein systematischer Vergleich beider Methoden bereitstellt werden. Zukünftige Studien sollten dennoch versuchen mögliche Übertragungseffekte zu minimieren.

Der Einsatz von verschiedenen Evaluatorenteams, wobei jedes Team nur eine der getesteten Evaluationsmethoden durchführt, hat sich als gute Maßnahme gegen Übertragungseffekte und den „Evaluatoreffekt“ im Allgemeinen erwiesen (Hornbæk & Frøkjær, 2008), war aber aufgrund der personellen und zeitlichen Beschränkungen im Rahmen des Projekts 3D-GUIde und damit dieser Arbeit nicht durchführbar. Dabei ist besonders spannend zu untersuchen, inwiefern Evaluatoren mit höherem Vorwissen (d.h. keine Studierenden, sondern Experten aus dem Feld) mit den Rhythmus-Peaks zurechtkommen und wie sich dabei Gründlichkeit, Gültigkeit, Zuverlässigkeit und zeitliche Anwendungseffizienz bei Berücksichtigung aller kritischen Ereignisse (d.h. weniger strikte Anwendung von IntuiBeat-F) und bei ausschließlicher Berücksichtigung mithilfe von Rhythmus-Peaks erkannter kritischer Ereignisse (d.h. strikte Anwendung von IntuiBeat-F) verändern.

Schließlich setzte sich die Stichprobe in allen vier Experimenten ausschließlich aus Studierenden der Medienkommunikation und Mensch-Computer-Systeme zusammen, die alle an der Julius-Maximilians-Universität Würzburg studieren. Es konnte dementsprechend insbesondere bei den CUIs (Experiment 4 und 6) eine sehr geringe Vorerfahrung bei der Nutzung von CUIs festgestellt werden. Auf diese Weise konnten zwar viele reale Nutzungsprobleme gefunden werden, wobei jedoch anzunehmen ist, dass Nutzer, die schon viel Vorerfahrung bei der Nutzung von CUIs mitbringen andere Nutzungsprobleme haben werden. Aufgrund der Tatsache, dass im Rahmen des Projekts 3D-GUIde Interaktionspatterns entwickelt werden sollten, die die von LaViola et al. (2017) beschriebenen domänenübergreifenden, grundlegenden Nutzeraufgaben Selektion, Manipulation, Navigation, Zeicheneingaben und Systemsteuerung unterstützen und diese somit auch von Novizen intuitiv nutzbare sein sollten, war die Verwendung einer technikaffinen Studierendenstichprobe schon angemessen. Das primäre Ziel dieser Arbeit eine formative Evaluationsmethode für intuitive Benutzung bereitzustellen, die im Rahmen des Projekts 3D-GUIde zur zeitlich effizienten Entwicklung von User Interfaces (speziell 3D-CUIs) genutzt werden kann, war mit diesen Stichproben somit erfüllbar. Zukünftige Studien sollten sich jedoch auch mit der Meta-Evaluation von IntuiBeat-F mit anderen Stichproben beschäftigen und dadurch auch die externe Gültigkeit von IntuiBeat-F zeigen.

Zusammenfassend kann festgehalten werden, dass trotz der oben genannten methodischen Limitationen und Alternativerklärungen mithilfe der vorliegenden vier Studien zum ersten Mal in der Forschungsliteratur eine andere Methode als ein Nutzertest für die formative Evaluation intuitiver Benutzung mit hoher zeitlicher Effizienz zum Einsatz kam und positive Ergebnisse seiner wissenschaftlicher Güte vorliegen. Mithilfe der Rhythmus-Peaks können Evaluatoren mit hoher Wahrscheinlichkeit auf, auf reale Nutzungsprobleme hinweisende, kritische Ereignisse aufmerksam gemacht werden, und sich damit nicht nur viel Zeit bei der Durchführung, sondern auch bei der Entscheidung über die Echtheit eines Nutzungsproblems sparen. Während man bei der Anwendung eines Nutzertests (egal welches Think-Aloud-Protokoll genau zum Einsatz kommt) noch zusätzlich untersuchen muss, welche Nutzungsprobleme nun speziell die intuitive Benutzung beeinträchtigen und für eine intuitiv benutzbare Software beseitigt werden müssen, bietet IntuiBeat-F dies als erste dedizierte formative Evaluationsmethode für intuitive Benutzung direkt an. Auf diese Weise konnten die damit abgeleiteten Erkenntnisse für eine Reihe von Interaktionspatterns im 3D-Bereich genutzt werden, die im Rahmen des Projekts 3D-GUIde entwickelt wurden (siehe Burmester et al., 2018). Dabei zeigten die vier Experimente außerdem, dass IntuiBeat-F sogar bei weniger strikter Anwendung ohne Verwendung einer zusätzlichen Analysesoftware (Experimente 4 und 5), die den Evaluator bei der strikteren Anwendung von IntuiBeat-F unterstützt, wissenschaftliche Güte und zeitliche Anwendungseffizienz attestiert werden kann. Um jedoch auf Nummer sicher zu gehen, sollten Praktiker auf die zusätzliche Analysesoftware zurückgreifen (Experimente 6 und 7) und ihnen damit durch die daraus folgende striktere Anwendung von IntuiBeat-F eine sowohl wissenschaftlich als auch zeitlich effizientere formative Evaluation intuitiver Benutzung ermöglicht werden.

Vergleicht man IntuiBeat-F und den Nutzertest mit retrospektivem Think-Aloud-Protokoll über alle vier Experimente hinweg schließlich noch auf qualitativer Ebene (d.h. welche Arten von Ursachen stecken hinter den gefundenen Nutzungsproblemen im Sinne der handlungsorientierten Fehlertaxonomie) und berücksichtigt dabei die damit verbundene hohe Beurteilerübereinstimmung bei der Klassifikation der Nutzungsprobleme ($\kappa \geq .63$), kann man keinen nennenswerten Unterschied zwischen den beiden Methoden feststellen. Dies ist nicht verwunderlich, wenn man den theoretischen Hintergrund von IntuiBeat-F berücksichtigt. Als eine anhand von Inhibitionsprozessen mentale Beanspruchung messende Methode erzeugt diese, laut dem Modell multipler Ressourcen von Wickens (2008), zwar die nötige Interferenz zur Erkennung von mentaler Beanspruchung verursachenden kritischen Ereignissen, ist dabei jedoch nicht besonders intrusiv. Demzufolge erhöht mit hoher Wahrscheinlichkeit IntuiBeat-F zwar die Sensitivität, indem es auch kritische Ereignisse durch Rhythmus-Peaks sichtbar macht, die vielleicht bei einem Nutzertest gar nicht aufgefallen wären (d.h. Rhythmus-Peaks lassen sich zumindest durch die verbesserte Darstellung in den letzten beiden Experimenten weniger leicht übersehen), beeinflusst aber nicht die Systeminteraktion an sich (d.h. ist intrusiv), weswegen die gefundenen Nutzungsprobleme auch nicht plötzlich andere Ursachen haben können. Aufgrund der Tatsache, dass Nutzungsprobleme erst durch die tatsächliche Interaktion eines Nutzers mit dem System zustande kommen, wäre eine derartige Änderung problematisch, da man nicht mehr von den mit IntuiBeat-F ermittelten Ergebnissen auf die „normale“ Systeminteraktion schließen kann, bei der der Nutzer keine zusätzliche explizite Zweitaufgabe in Form einer Rhythmusaufgabe durchführen muss. Wie bei einem gewöhnlichen Nutzertest mit retrospektivem Think-Aloud-Protokoll, in dem der Nutzer so natürlich wie in einer Laborsituation

möglich, seine Systeminteraktion durchführt, gestattet es IntuiBeat-F einem Evaluator, speziell intuitive Benutzung aufgrund der bereitgestellten Rhythmus-Peaks gründlicher, gültiger und zeitlich effizienter formativ zu evaluieren.

7 Zusammenfassende Diskussion und Ausblick

Die zentrale Zielsetzung dieser Dissertation besteht darin herauszufinden, wie die intuitive Benutzung von User Interfaces sowohl formativ als auch summativ mit möglichst hoher zeitlicher Anwendungseffizienz evaluiert werden kann, und inwiefern die vorgestellte Methode *IntuiBeat* (d.h. IntuiBeat-F und IntuiBeat-S) dazu geeignet ist. Um eine Antwort auf diese zentrale Fragestellung liefern zu können, ist es nun an der Zeit die Inhalte der vorherigen Kapitel zu verbinden und die dort stattgefundene Diskussion bezüglich der Evaluation intuitiver Benutzung auf Basis von Rhythmusabweichungen zusammenzufassen. Hierzu gibt dieses abschließende Kapitel zunächst eine Übersicht der Zielsetzung dieser Arbeit und diskutiert anschließend, welche Arten von wissenschaftlichen Beiträgen durch diese Arbeit für den HCI-Bereich geliefert werden können, welche Limitationen damit in Verbindung stehen, und an welchen Punkten zukünftige Forschung anknüpfen kann. Das Fazit dieser Arbeit ist, dass *IntuiBeat* für die Evaluation intuitiver Benutzung mit hoher zeitlicher Anwendungseffizienz eingesetzt werden kann.

7.1 Zielsetzung der Arbeit: Entwicklung einer neuen Methode zur Evaluation intuitiver Benutzung

Im Rahmen dieser Arbeit sollte mit IntuiBeat eine Evaluationsmethode für intuitive Benutzung vorgestellt und deren Güte (d.h. wissenschaftliche Güte und zeitliche Anwendungseffizienz) in Form einer Meta-Evaluation bewertet werden. Die Anforderung an eine hohe zeitliche Anwendungseffizienz ergab sich aus dem Anwenderprojekt 3D-GUIde, welches als Rahmenbedingung für die Meta-Evaluation der formativen (d.h. IntuiBeat-F) und der summativen (d.h. IntuiBeat-S) Evaluationsmethode fungierte (siehe Kapitel 1). Dagegen mangelt es existierenden Benchmarks im formativen (d.h. Nutzertest mit retrospektivem Think-Aloud-Protokoll, siehe Abschnitt 3.5) und summativen (d.h. CHAI-Methode, siehe Abschnitt 3.6) Bereich an zeitlicher Anwendungseffizienz. So besitzt der formative Benchmark eine geringe zeitliche Anwendungseffizienz wegen des retrospektiven Interviews, bei dem sich der Evaluator zusammen mit dem Nutzer dessen Systemnutzung im Nachgang auf Video anschaut und dann mithilfe eines Think-Aloud-Protokolls ohne zusätzliche Unterstützung Nutzungsprobleme identifiziert. Auch der summative Benchmark besitzt aufgrund der damit verbundenen Videoanalyse, bei der jeder Nutzerklick im Hinblick auf intuitive Benutzung mithilfe des CHAI-Bewertungsschemas von einem Experten beurteilt werden muss, eine geringe zeitliche Anwendungseffizienz.

Daher wurde im Rahmen dieser Dissertation die Methode IntuiBeat auf Basis der Rhythmusmethode von Park und Brünken (2015) entwickelt, welche die intuitive Benutzung formativ (d.h. IntuiBeat-F) und summativ (d.h. IntuiBeat-S) basierend auf Rhythmusabweichungen (d.h. Abweichungen von einer zuvor ermittelten Rhythmus-Baseline) evaluiert. Der Nutzer muss hierzu während der Systemnutzung eine Rhythmuszweitaufgabe ausführen, bei der er einen zuvor erlernten Rhythmus mit dem Fuß klopfen muss, der das bewusste

Einhalten von Pausen erfordert (siehe Abschnitt 2.2). Hierfür werden mental beanspruchende Inhibitionsprozesse der zentralen Exekutive des Arbeitsgedächtnisses benötigt, um bewusst die Pausen des gelernten Rhythmus einhalten zu können (Korbach et al., 2018; Park & Brünken, 2015).

Um Rhythmusabweichungen mit intuitiver Benutzung in Verbindung zu bringen, wurde hierfür zunächst in Kapitel 2 eine theoretische Grundlage geschaffen, indem, mithilfe von Default-Interventionist-Zwei-Prozess-Theorien, die für die Evaluation intuitiver Benutzung zentralen drei Charaktermerkmale aus allen mit intuitiver Benutzung assoziierten Merkmalen ausgewählt und diese in einer Messdefinition festgehalten wurden. Laut dieser Messdefinition (siehe Teilabschnitt 2.2.2) lässt sich intuitive Benutzung als das Ausmaß definieren, mit dem ein Produkt mental effizient (d.h. objektives zentrales Charaktermerkmal) und effektiv (d.h. pragmatisches zentrales Charaktermerkmal) genutzt wird. Dies ist verbunden mit einem starken metakognitiven Gefühl von Flüssigkeit (d.h. subjektives zentrales Charaktermerkmal). Intuitive Benutzung lässt sich dabei objektiv anhand der mentalen Beanspruchung bei der kognitiven Informationsverarbeitung im Arbeitsgedächtnis und damit verbundenen objektiven Korrelaten messen. Objektive Korrelate sind messbare Merkmale intuitiver Benutzung, die im Gegensatz zum objektiven zentralen Charaktermerkmal nicht notwendigerweise immer alle gleichzeitig auftreten müssen, denen aber die mentale Beanspruchung implizit zugrunde liegt und ihre Ausprägung bestimmt.

Ferner lässt sich intuitive Benutzung auch subjektiv anhand des Gefühls von Flüssigkeit und damit verbundenen subjektiven Korrelaten beurteilen. Subjektive Korrelate bezeichnen subjektiv messbare Merkmale intuitiver Benutzung, die im Gegensatz zum subjektiven zentralen Charaktermerkmal nicht notwendigerweise immer alle gleichzeitig auftreten müssen, denen aber implizit das metakognitive Gefühl von Flüssigkeit zugrunde liegt und ihre Ausprägung bestimmt.Da Inhibitionsprozesse als gute Indikatoren für Prozesse der zentralen Exekutive des Arbeitsgedächtnisses und damit der gesamten mentalen Beanspruchung des Nutzers gelten (J. D. Cohen et al., 1997; A. Miyake et al., 2000; Stanovich et al., 2014), kann eine Inhibition erfordernde Rhythmuszweitaufgabe und die damit verbundene Methodik (d.h. IntuiBeat-F und IntuiBeat-S) potentiell einen neuen Benchmark für die formative und summative Evaluation intuitiver Benutzung darstellen. Jedoch existieren in der Forschungsliteratur eine Reihe von ungeklärten Aspekten und Limitationen vorheriger Experimente, die sich mit der formativen und summativen Evaluation von mentaler Beanspruchung auf Basis einer Rhythmuszweitaufgabe beschäftigen (siehe Kapitel 4). Diese Dissertation widmet sich daher den damit verbundenen Forschungslücken anhand der Beantwortung von drei Forschungsfragen.

Zum einen konnten bisherige Arbeiten noch keinen vollständigen empirischen Nachweis bezüglich der Höhe der wissenschaftlichen Güte einer auf Rhythmusabweichungen basierenden summativen Evaluationsmethode auf Basis der drei summativen Hauptgütekriterien *Objektivität*, *Reliabilität* und *Validität* liefern (Forschungsfrage 1). Zum anderen konnten bisherige Arbeiten auch noch keinen vollständigen empirischen Nachweis bezüglich der Höhe der wissenschaftlichen Güte einer auf Rhythmusabweichungen basierenden formativen Evaluationsmethode auf Basis der drei formativen Hauptgütekriterien *Gründlichkeit*, *Gültigkeit* und *Zuverlässigkeit* liefern (Forschungsfrage 2). Schließlich wurde sowohl im Bereich der formativen Evaluation als auch im Bereich der summativen Evaluation intuitiver Benutzung noch nicht untersucht, wie hoch die zeitliche Anwendungseffizienz von

IntuiBeat-S und IntuiBeat-F im Vergleich zu bereits vorhandenen Evaluationsmethoden (d.h. aktuellen Benchmarks) für intuitive Benutzung ist (Forschungsfrage 3).

Durch die Beantwortung dieser drei Forschungsfragen und durch die damit verbundene Vorarbeit leistet diese Dissertation für die HCI verschiedene Beiträge. Laut Wobbrock und Kientz (2016) kann ein Forschungsbeitrag theoretisch, empirisch oder methodisch sein, es sich um einen Überblick des Forschungsfeldes (d.h. Übersichtsarbeit bzw. Review), einen Datensatz oder eine Meinung handeln, sowie darin ein neues Werkzeug beschrieben sein. Im Rahmen dieser Dissertation wurde ein theoretischer, methodischer und empirischer Forschungsbeitrag sowie ein Forschungsbeitrag im Sinne eines neuen Werkzeugs und eines Reviews (d.h. Übersichtsarbeit) geleistet. Die folgenden fünf Teilabschnitte sollen nun detailliert auf die konkreten Forschungsbeiträge eingehen und definieren jeweils einleitend die Art des Forschungsbeitrags.

7.2 Übersichtsarbeiten

Übersichtsarbeiten oder Reviews stellen in der HCI Versuche dar, vorhandene Forschungsarbeiten im Feld zu überprüfen und miteinander zu verknüpfen. Ziel ist dabei das Erkennen von Trends, aktuellen Themen und Lücken in der Forschungsliteratur (Wobbrock & Kientz, 2016). Als eine qualitative Zusammenfassung der Ergebnisse einzelner Experimente (Blettner, Sauerbrei, Schlehofer, Scheuchenpflug, & Friedenreich, 1997) konzentrieren sich Übersichtsarbeiten im Vergleich zu den anderen, von Wobbrock und Kientz (2016) genannten Beitragsarten auf die Vergangenheit, um die verfügbare Literatur des Forschungsfeldes zu strukturieren und über deren aktuelle Bedeutung für die HCI zu reflektieren. Laut Wobbrock und Kientz (2016) erscheinen Übersichtsarbeiten üblicherweise dann, wenn das Forschungsfeld eine gewisse Reife erreicht hat. Es ist nicht ungewöhnlich, dass Übersichtsarbeiten verglichen mit den anderen Beitragsarten sehr umfangreich sind und manchmal hunderte von wissenschaftlichen Referenzen enthalten. Dabei ist jedoch laut Wobbrock und Kientz (2016) entscheidend, dass Übersichtsarbeiten nicht nur reine Auflistungen von Forschungsarbeiten darstellen, sondern vielmehr die Überprüfung und Synthese vorhandener Literatur in den Vordergrund rücken. Die vorliegende Dissertation leistet hauptsächlich zwei verschiedene Übersichten, die im Folgenden zusammenfassend vorgestellt und auch hinsichtlich zukünftiger Forschungsmöglichkeiten diskutiert werden sollen.

7.2.1 Übersicht über Definitionen intuitiver Benutzung

Bevor die zentrale Frage dieser Dissertation, ob sich die intuitive Benutzung von User Interfaces (speziell 3D-CUIs) sowohl formativ als auch summativ mit möglichst hoher zeitlicher Anwendungseffizienz evaluieren lässt, beantwortet werden kann, wurden in Abschnitt 2.1 dieser Dissertation zunächst alle drei im Forschungsfeld zu intuitiver Benutzung vorhandenen Hauptdefinitionen rezensiert und bezüglich der darin enthaltenen, intuitive Benutzung charakterisierenden Merkmale verglichen, sowie Unterschiede zwischen den drei, für die Definitionen verantwortlichen Forschergruppen (d.h. IUUI-, QUT-, und INTUI-Forschergruppe) herausgearbeitet.

Die Definitionen der IUUI- und QUT-Forschergruppe konzentrieren sich dabei vorwiegend auf objektiv beurteilbare Merkmale intuitiver Benutzung, wobei die IUUI-Forschergruppe (siehe Hurtienne, 2011; Mohs, Hurtienne, Kindsmüller et al., 2006; Naumann et al., 2007) die hohe mentale Effizienz und Effektivität bei der Nutzung und die QUT-Forschergruppe (siehe Blackler, 2006, 2018; Blackler et al., 2010) die hohe zeitliche Effizienz während der Nutzung und die geringe retrospektive Verbalisierbarkeit der Systemnutzung in den Fokus rücken (siehe Abschnitt 2.1). Im Gegensatz zu diesen beiden Definitionen fokussieren sich die Arbeiten der INTUI-Forschergruppe (siehe Diefenbach & Ullrich, 2015; Ullrich, 2013, 2014), die keine explizite Definition beinhalten, auf vorwiegend subjektiv beurteilbare Merkmale intuitiver Benutzung, wie die geringe wahrgenommene Mühelosigkeit, das starke wahrgenommene Bauchgefühl, die geringe wahrgenommene retrospektive Verbalisierbarkeit und das stark wahrgenommene magische Erleben.

Neben den unterschiedlichen Schwerpunkten, die die Arbeiten der drei Forschergruppen besitzen, wurde in Abschnitt 2.1 zudem explizit für jede Forschergruppe eine Liste von Charaktermerkmalen herausgearbeitet, die mit intuitiver Benutzung assoziierten Merkmale nach Forschergruppe gruppiert und in Form einer zusammenfassenden Liste aller, mit intuitiver Benutzung assoziierten Charaktermerkmale bereitgestellt (siehe Teilabschnitt 2.1.4). Im Zuge dessen wurde ferner verdeutlicht, dass nicht alle mit intuitiver Benutzung assoziierten Merkmale von allen Forschergruppen gleichermaßen zur Charakterisierung des Konstrukts genutzt werden. Des Weiteren konnte im Rahmen dieser kritischen Betrachtung gezeigt werden, dass laut der Definitionen der IUUI- (siehe Naumann et al., 2007) und QUT-Forschergruppe (siehe Blackler, 2006) die mit intuitiver Benutzung assoziierten Merkmale immer gleichzeitig in gleicher Ausprägung auftreten müssen, und sie so das Ausmaß intuitiver Benutzung auf einem Kontinuum beschreiben. Die INTUI-Forschergruppe (siehe Ullrich, 2014), die sich von den anderen Forschergruppen mit ihrer eingenommenen Erlebnisperspektive bereits generell unterscheidet, hat auch hier eine andere Auffassung und argumentiert, dass sich das Ausmaß intuitiver Benutzung nicht notwendigerweise immer gleichermaßen in allen vorgeschlagenen Merkmalen widerspiegeln muss. Stattdessen bestimmt die relative Ausprägung der verschiedenen Merkmale zueinander (d.h. INTUI-Patterns, siehe Ullrich, 2014) über das Ausmaß intuitiver Benutzung. Im Gegensatz zu den anderen beiden Forschergruppen macht die INTUI-Forschergruppe diese Auffassung nicht explizit in Form einer Definition intuitiver Benutzung deutlich, sondern beschreibt stattdessen intuitive Benutzung direkt anhand von vier Komponenten bzw. vier Merkmalen (siehe Ullrich, 2014).

Im Vergleich zu aktuell verfügbaren Reviews (z.B. Blackler, 2018; Blackler & Hurtienne, 2007; Blackler & Popovic, 2015) besitzt das im Rahmen dieser Arbeit geleistete Review, durch die Betrachtung des Konstrukts aus den unterschiedlichen Perspektiven der drei genannten Hauptforschergruppen und durch dessen Charakterisierung anhand der damit assoziierten Merkmale, eine hohe Strukturiertheit, wie sie von Wobbrock und Kientz (2016) gefordert wird. Frühere Reviews betrachten zwar ebenfalls intuitive Benutzung aus der Perspektive verschiedener Forschergruppen (siehe Blackler, 2018; Blackler & Hurtienne, 2007; Blackler & Popovic, 2015), wobei nicht einmal das aktuellste Review (siehe Blackler, 2018) alle mit intuitiver Benutzung assoziierten Merkmale herausarbeitet, die dafür verantwortlichen Forschergruppen explizit benennt und die Charaktermerkmale in Form einer Übersichtsliste, wie in Teilabschnitt 2.1.4 dieser Dissertation geschehen, bereitstellt. Die im Rahmen von Dissertationen erbrachten grundlegenden Übersichtsarbeiten der einzel-

nen Forschergruppen (siehe Blackler, 2006; Hurtienne, 2011; Ullrich, 2014) konnten zwar jeweils eine Reihe von mit intuitiver Benutzung assoziierten Merkmalen herausarbeiten, diese aber aufgrund ihrer zeitlichen Versetzung (d.h. es liegen mehrere Jahre zwischen den Arbeiten) noch nicht vollständig mit von anderen Forschergruppen assoziierten Merkmalen in Verbindung bringen.

Damit ergänzt die im Rahmen dieser Dissertation erbrachte *Übersicht vorhandener Definitionen intuitiver Benutzung* frühere Übersichtsarbeiten. Aufgrund der höheren Reife des Forschungsfeldes zum Zeitpunkt ihrer Anfertigung war dieser eine differenziertere Sichtweise auf das Konstrukt *intuitive Benutzung*, ein hohes Maß an Strukturiertheit und eine tiefe Synthese möglich. So wurde bei der Extraktion der mit intuitiver Benutzung assoziierten Charaktermerkmale stets darauf geachtet, (1) die Terminologie zu vereinheitlichen (d.h. unter Berücksichtigung der zugrunde liegenden Literatur wurde kritisch reflektiert, inwiefern es sich bei zwei Charaktermerkmalen um das gleiche Charaktermerkmal handelt), (2) zu verdeutlichen, inwiefern Merkmale zusammen gleichermaßen das Ausmaß intuitiver Benutzung beschreiben oder ob dieses Ausmaß durch die relative Ausprägung der Merkmale bestimmt wird, und (3) die hinter diesen Merkmalen stehenden theoretischen Grundlagen durch eine metatheoretische Auseinandersetzung mithilfe der Handlungsregulationstheorie (siehe Hacker, 1986) und Zwei-Prozess-Theorien (siehe Evans & Stanovich, 2013) zusammenzuführen. Letzterer Aspekt soll im folgenden Abschnitt diskutiert werden.

Aus der Beurteilung der in Abschnitt 2.1 beschriebenen Übersicht über vorhandene Definitionen intuitiver Benutzung ergeben sich auch eine Reihe von Themen, mit denen sich zukünftige Forschung auseinandersetzen kann. Es lässt sich durch die starke Fokussierung auf die drei Hauptforschergruppen nicht verhindern, dass von diesen Forschergruppen unabhängige Forscher (z.B. O'Brien et al., 2010; Reinhardt & Hurtienne, 2018) weniger stark repräsentiert sind und zukünftige Reviews daher noch in puncto Fairness (siehe Wobbrock & Kientz, 2016) optimiert werden können. Des Weiteren muss im Zuge der Vereinheitlichung der Terminologie, der mit intuitiver Benutzung assoziierten Merkmalen eingeräumt werden, dass insbesondere die vier Merkmale, Mühelosigkeit, Verbalisierbarkeit, Bauchgefühl und magisches Erleben einen gewissen Interpretationsspielraum zuließen. So wurde basierend auf den Arbeiten der INTUI-Forschergruppe (siehe Ullrich, 2014) und der dort referenzierten Forschungsliteratur beispielsweise Mühelosigkeit mit den von den anderen beiden Forschergruppen vorgeschlagenen Merkmalen der zeitlichen und mentalen Effizienz in Verbindung gebracht. Des Weiteren wurde bei ursprünglich als lediglich subjektiv beurteilbaren, von der INTUI-Forschergruppe vorgeschlagenen Merkmalen auch deren objektive Beurteilbarkeit implizit angenommen, da diese Merkmale bereits von den anderen beiden Forschergruppen als objektiv bewertbar eingeordnet wurden (siehe Abschnitt 2.1). Zukünftige Reviews könnten hier konservativer bleiben und nicht die Synthese der Arbeiten in den Vordergrund stellen, sondern sich eher auf die Optimierung der Fairness des Reviews konzentrieren. Dies würde bedeuten die Position der referenzierten Arbeiten möglichst originalgetreu wiederzugeben (siehe Wobbrock & Kientz, 2016) und den Interpretationsspielraum nicht auszuschöpfen.

Schließlich lässt sich die von Wobbrock und Kientz (2016) geforderte Fairness, sowie andere für Reviews wichtige Beurteilungskriterien (z.B. Vollständigkeit, Gründlichkeit, Strukturiertheit, Tiefe der Synthese) dadurch verbessern, das Review nicht wie in dieser Arbeit

narrativ basierend auf bereits vorhandenen Reviews (z.B. Blackler, 2018; Blackler & Hurtienne, 2007; Blackler & Popovic, 2015) und darin referenzierten Arbeiten zu erstellen, sondern noch zusätzlich ein davon unabhängiges Review vorzunehmen. Ein narratives Review gibt einen breiten Überblick über das untersuchte Forschungsfeld (Baumeister & Leary, 1997; Blettner et al., 1997; Derish & Annesley, 2011; Ressing, Blettner, & Klug, 2009), die Auswahl der der berücksichtigten Artikel erfolgt jedoch subjektiv (z.B. nur bestimmte Arbeiten werden berücksichtigt, siehe Lodge, 1981) und kann dadurch die Fairness verzerren. Daher sollten zukünftig auch systematische Reviews im Forschungsfeld zu intuitiver Benutzung vorgenommen werden. Systematische Reviews erheben laut Derish und Annesley (2011) den Anspruch, unter Berücksichtigung von vor dem Review definierten Ein- und Ausschlusskriterien möglichst alle zu einem bestimmten Thema veröffentlichten Arbeiten zu berücksichtigen (Derish & Annesley, 2011; Montori, Swiontkowski, & Cook, 2003; Ressing et al., 2009). Den Schwerpunkt eines systematischen Reviews bildet hierbei die systematische, kritische Extraktion relevanter Informationen aus den betrachteten Arbeiten.

7.2.2 Übersicht über Evaluationsmethoden intuitiver Benutzung

Neben der Übersicht über vorhandene Definitionen intuitiver Benutzung wurde noch eine zweite Übersichtsarbeit im Rahmen dieser Dissertation geleistet. Es wurde hier zum einen eine *Übersicht über vorhandene Methoden zur formativen Evaluation intuitiver Benutzung* (siehe Abschnitt 3.5) und zum anderen eine *Übersicht über vorhandene Methoden zur summativen Evaluation intuitiver Benutzung* (siehe Abschnitt 3.6) gegeben.

Im Zuge der *Übersicht über vorhandene formative Evaluationsmethoden* konnte unter Berücksichtigung aktueller Reviews (siehe Blackler et al., 2018) und unabhängiger Recherche auf Basis der in Teilabschnitt 2.2.2 abgeleiteten Messdefinition für intuitive Benutzung festgestellt werden, dass keine dedizierte formative Evaluationmethode für intuitive Benutzung existiert. Stattdessen adaptiert das Forschungsfeld einen Nutzertest mit Think-Aloud-Protokoll (d.h. entweder parallel, retrospektiv oder als Co-Discovery ausgeführt) aus dem allgemeinen Usability-Bereich für diesen Zweck (siehe Abschnitt 3.5). Hierbei wurde auch herausgearbeitet, dass das retrospektive Think-Aloud-Protokoll gegenüber den anderen beiden vorgestellten Varianten des „lauten Denkens“ (d.h. paralleles Think-Aloud-Protokoll, Co-Discovery) aufgrund seiner im Vergleich geringeren zeitlichen Anwendungseffizienz im Forschungsfeld zwar weniger verbreitet ist (siehe Reinhardt et al., 2018), aber gegenüber den anderen beiden Varianten eine höhere wissenschaftliche Güte aufweisen kann (siehe V. A. Bowers & Snyder, 1990; Hertzum et al., 2009; Kuusela & Pallab, 2000; Van den Haak & De Jong, 2003; Van Den Haak et al., 2003; Van den Haak et al., 2004). Der Nutzertest mit retrospektivem Think-Aloud-Protokoll wurde somit auch als aktueller Benchmark zur formativen Evaluation intuitiver Benutzung identifiziert. Er stellt damit auch das einzige verfügbare Quasi-Außenkriterium für eine Meta-Evaluation von IntuiBeat-F dar. Da der Nutzertest mit retrospektivem Think-Aloud-Protokoll wegen fehlender Unterstützung bei der retrospektiven Analyse der gefundenen kritischen Ereignisse eine niedrige zeitliche Anwendungseffizienz besitzt, wurde der Bedarf nach einer dedizierten formativen Evaluationsmethode für intuitive Benutzung explizit aufgezeigt. Dies rechtfertigte wiederum die Entwicklung und Meta-Evaluation von IntuiBeat-F im

Rahmen dieser Dissertation. Wegen des expliziten Hinweises auf den Bedarf an einer neuen formativen Evaluationsmethode unter Berücksichtigung der formalen, wissenschaftlichen, formativen Hauptgütekriterien (d.h. Gründlichkeit, Gültigkeit und Zuverlässigkeit) sowie der zeitlichen Anwendungseffizienz kann die im Rahmen des vorliegenden Reviews vorgenommene tiefe Synthese augenscheinlich als gründlich, vollständig und strukturiert eingeschätzt werden (siehe Wobbrock & Kientz, 2016).

Im Zuge der *Übersicht über vorhandene summative Evaluationsmethoden* konnte unter Berücksichtigung aktueller Reviews (siehe Blackler et al., 2018) und unabhängiger Recherche auf Basis der in Teilabschnitt 2.2.2 abgeleiteten Messdefinition für intuitive Benutzung festgestellt werden, dass im Forschungsfeld zwischen subjektiven und objektiven Methoden zur summativen Evaluation intuitiver Benutzung unterschieden wird (siehe Abschnitt 3.6). Als subjektive Methoden wurden der TFQ, der QUESI, der INTUI-Fragebogen, der SMEQ bzw. die SEA-Skala und der NASA-RTLX als die am weitest verbreiteten Maße vorgestellt. Der QUESI, der NASA-RTLX und der SMEQ bzw. die SEA-Skala konnten als geeignete Quasi-Außenkriterien für eine Meta-Evaluation von IntuiBeat-S identifiziert werden (siehe Teilabschnitt 3.6.1). Die letzten beiden Fragebögen, die ursprünglich aus der Forschung zu mentaler Beanspruchung und nicht direkt aus der Forschung zu intuitiver Benutzung stammen, wurden bereits in aktuellen Reviews (siehe Blackler et al., 2018) berücksichtigt, was nicht verwunderlich ist, da basierend auf der in dieser Arbeit vorgeschlagenen Messdefinition (siehe Teilabschnitt 2.2.2) und anderen Definitionen (siehe Teilabschnitt 2.1.4), die mentale Beanspruchung das zentrale objektive Merkmal zur Beurteilung des Ausmaßes an intuitiver Benutzung bildet.

Da subjektive Methoden aufgrund ihrer retrospektiven Erfassung unter eine Reihe von methodischen Limitationen, wie beispielsweise der Anfälligkeit gegenüber Primacy-Recency-Effekten (siehe Blackler, 2006), leiden, sollen laut Blackler et al. (2018) zusätzlich immer objektive Methoden zum Einsatz kommen. Im Rahmen des Teilabschnitts 3.6.2 wurden hierzu Verhaltensmaße, physiologische Maße, Hauptaufgabenleistungsmaße und Zweitaufgabenleistungsmaße vorgestellt. Es wurde dabei detailliert herausgearbeitet, warum die CHAI-Methode (siehe Reinhardt et al., 2018) als Verhaltensmaß den aktuellen Benchmark zur summativen Evaluation intuitiver Benutzung darstellt. Einen Grund für diese Entscheidung bildet die Tatsache, dass die wissenschaftliche Güte von physiologischen Maßen und der Q-Methode (siehe Asikhia, 2015), auf Basis der Hauptgütekriterien *Objektivität*, *Reliabilität* und *Validität* nicht ausreichend nachgewiesen ist. Des Weiteren weisen physiologische Maße unter anderem eine hohe Anfälligkeit gegenüber äußeren Umgebungseinflüssen (z.B. Auswirkung der Lichtverhältnisse bei okularen Maßen) auf. Zudem besteht die Problematik, dass physiologische Reaktionen des Körpers nicht notwendigerweise von einer einzigen Quelle verursacht werden (z.B. erhöhte Herzfrequenz bildet nicht notwendigerweise nur die erhöhte mentale Beanspruchung durch die Systemnutzung ab), was das Signal-Rausch-Verhältnis und damit die wissenschaftliche Güte solcher Maße negativ beeinflussen kann (Cowley et al., 2016). Ein weiterer Grund, der für die CHAI-Methode als summativer Benchmark spricht, ist, dass sich ein Hauptaufgabenleistungsmaß praktisch nur in Kombination mit einem anderen objektiven Maß einsetzen lässt (siehe Cain, 2007) und so keine Rückschlüsse auf die intuitive Benutzung während der eigentlichen Systemnutzung zulässt (z.B. die Anzahl korrekt abgeschlossener Aufgaben erlaubt kein Urteil über die intuitive Benutzung einzelner Interaktionen mit einem bestimmten Interaktionselement).

Bezüglich der vorgestellten Zweitaufgabenleistungsmaße zur summativen Evaluation intuitiver Benutzung wurde basierend auf dem Modell multipler Ressourcen von Wickens (2008) herausgearbeitet, dass es schwer möglich ist, eine universelle Zweitaufgabe zu wählen. Dies liegt daran, dass eine universelle Zweitaufgabe möglichst vollständig die intuitive Benutzung anhand der mentalen Beanspruchung abbilden und dabei nicht nur Teilaspekte (z.B. modalitätsspezifische mentale Beanspruchung aufgrund visueller Belastung) berücksichtigen soll, ohne dabei gleichzeitig eine hohe Intrusion (z.B. ablenken) durch die Zweitaufgabe selbst in Kauf zu nehmen, da die Überschneidung der von Haupt- und Zweitaufgabe benötigten Ressourcen darüber entscheidet, wie viele Teilaspekte mentaler Beanspruchung durch die Messung erfasst werden können. Im Zuge eines Resümees wurde an dieser Stelle begründet (siehe Teilabschnitt 3.6.2.4), dass aktuelle Zweitaufgaben (siehe Gawron, 2019) diese Anforderung nicht erfüllen können. Stattdessen könnte eine auf Inhibition basierende Zweitaufgabe diese methodische Problematik umgehen (Park & Brünken, 2015) und sowohl zur formativen als auch zur summativen Evaluation intuitiver Benutzung eingesetzt werden. Im Vergleich zum aktuellen summativen Benchmark, der CHAI-Methode, die aufgrund der damit verbundenen aufwendigen Videoanalyse eine geringe zeitliche Anwendungseffizienz aufweist, wurde an dieser Stelle aufgezeigt, inwiefern die Entwicklung und Meta-Evaluation von IntuiBeat-S im Rahmen dieser Dissertation gerechtfertigt werden kann.

Im Vergleich zu früheren Reviews (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) fokussiert sich das im Rahmen dieser Arbeit geleistete Review ausschließlich auf Evaluationsmethoden. Reviews wie das von Blackler und Hurtienne (2007) beleuchteten Evaluationsmethoden und Gestaltungsmethoden gleichermaßen, weswegen eine vertiefte Auseinandersetzung mit Evaluationsmethoden hier nicht erfolgen konnte. Aktuellere Reviews (d.h. Blackler & Popovic, 2015; Blackler et al., 2018) leisten durch ihren stärkeren Fokus auf Evaluationsmethoden zwar eine tiefere Auseinandersetzung, jedoch erfolgt in beiden Reviews keine Bewertung der wissenschaftlichen Güte der summativen Evaluationsmethoden bezüglich der formalen summativen Hauptgütekriterien. Wenn die Güte der Evaluationsmethode doch einmal Thema ist, erfolgt meist nur eine Bewertung der Methoden unter Berücksichtigung nicht formaler Gütekriterien (z.B. Sensitivität, Diagnostizität) ohne Verwendung eines Außenkriteriums (siehe Blackler et al., 2018). Das im Rahmen dieser Dissertation geleistete Review lieferte stattdessen eine formale Bewertung der wissenschaftlichen Güte, sowie der zeitlichen Anwendungseffizienz als praxisrelevanten wichtigen Teilaspekt der praktischen Güte (siehe Schmidt-Kretschmer & Blessing, 2006) der im Forschungsfeld zu intuitiver Benutzung am weitesten verbreiteten Evaluationsmethoden (siehe Teilabschnitt 3.6.2).

Des Weiteren konnten im Rahmen dieser Dissertation aufgrund der Berücksichtigung der in Teilabschnitt 2.2.2 beschriebenen Messdefinition und der Fokussierung auf mentale Beanspruchung als das objektive Merkmal zur Beurteilung des Ausmaßes intuitiver Benutzung auch insbesondere physiologische Maße und Zweitaufgabenleistungsmaße ausgiebig berücksichtigt werden. Beide Arten werden in aktuellen Reviews (siehe Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) entweder überhaupt nicht oder in den grundlegenden Dissertationen der jeweiligen Forschergruppen (siehe Blackler, 2006; Hurtienne, 2011; Ullrich, 2014) nur am Rande diskutiert. Auch dann findet keinerlei Bewertung der wissenschaftlichen Güte der Maße bezogen auf die summative Evaluation intuitiver Benutzung statt. Zusammenfassend erfolgte die im Rahmen dieser Dissertation

geleistete Übersicht über vorhandene summative Evaluationsmethoden möglichst vollständig, gründlich, strukturiert und fair mit einer tiefen Synthese (Wobbrock & Kientz, 2016), weswegen dieses Review eine wertvolle Ergänzung bisheriger Reviews (z.B. Blackler & Hurtienne, 2007; Blackler & Popovic, 2015; Blackler et al., 2018) darstellen kann.

Trotzdem ergibt sich sowohl bezüglich der Übersicht über vorhandene formative Evaluationsmethoden (siehe Abschnitt 3.5) als auch bezüglich der Übersicht über vorhandene summative Evaluationsmethoden (siehe Abschnitt 3.6) eine Reihe von Verbesserungsmöglichkeiten, die durch zukünftige Forschung realisiert werden können. Wie bereits in der vorherigen Übersichtsarbeit beschrieben, erfolgten beide Übersichten über Evaluationsmethoden ebenfalls in rein narrativer Form, weswegen systematische Reviews eine sinnvolle Ergänzung darstellen würden (siehe Derish & Annesley, 2011). Außerdem betrachteten beide Reviews neben den wissenschaftlichen formalen Hauptgütekriterien lediglich die zeitliche Anwendungseffizienz als Nebengütekriterium. Zukünftige Reviews könnten herausarbeiten, inwiefern andere Nebengütekriterien aus dem formativen (z.B. Downstream Utility) und dem summativen (z.B. Normierung) Bereich von Evaluationsmethoden für intuitive Benutzung bereits erfüllt werden. Schließlich könnten zukünftige Reviews noch genauer beleuchten, in welchen Einsatzgebieten bzw. in welchen genauen Nutzungskontexten die einzelnen im Forschungsfeld verfügbaren Evaluationsmethoden bereits eingesetzt und formal evaluiert wurden. Auf diese Weise kann mit der in Teilabschnitt 3.2.5 angesprochenen Problematik der Übergeneralisierung von Methoden in der HCI im Forschungsfeld zu intuitiver Benutzung umgegangen werden. Da sich diese Dissertation speziell auf 3D-CUIs als Anwendungsdomäne fokussiert hat, konnte eine derart vertiefte Auseinandersetzung im Rahmen dieser Dissertation jedoch nicht erfolgen.

7.3 Theoretischer Forschungsbeitrag

Laut Wobbrock und Kientz (2016) umfassen theoretische Forschungsbeiträge in der HCI neue Modelle, Prinzipien, Definitionen, Konzepte, Frameworks oder wichtige Veränderungen davon. Dabei können theoretische Forschungsbeiträge sowohl quantitativ als auch qualitativ, sowie deskriptiv als auch prädiktiv sein. Sie sind in allen Fällen jedoch so strukturiert, dass sie für zukünftige Forschung inhärent nutzbare Theorien zur Verfügung stellen und so die „Art des Denkens“ basierend auf diesem Verständnis fördern können. Die Beurteilung eines theoretischen Forschungsbeitrags erfolgt üblicherweise durch empirische Beobachtung, weswegen dadurch zeitgleich auch ein empirischer Forschungsbeitrag geleistet wird (Wobbrock & Kientz, 2016).

Während der Erstellung der *Übersicht über vorhandene Definitionen intuitiver Benutzung* in Abschnitt 2.1 wurde festgestellt, dass sich die drei Hauptforschergruppen (d.h. IUUI, QUT und INTUI) nicht einig sind, (1) anhand welcher Merkmale sich das Ausmaß intuitiver Benutzung beurteilen lässt und (2) ob diese Merkmale dafür alle gleichzeitig, sowie in gleicher Ausprägung (siehe IUUI- und QUT-Forschergruppe in den Teilabschnitten 2.1.1 und 2.1.2) oder nur bestimmte Merkmale in relativer Ausprägung zueinander (siehe INTUI-Forschergruppe in Teilabschnitt 2.1.3) vorliegen müssen. Wie zu Beginn des Abschnitts 2.2 erläutert, ist hierfür, wie auch in der allgemeinen Intuitionsforschung außerhalb der HCI, deren unterschiedliche wissenschaftliche Grundlagen und die Verortung

intuitiven Handelns innerhalb des menschlichen Denkens verantwortlich (siehe Zander et al., 2016). Als theoretischer Forschungsbeitrag dieser Dissertation wurden die Arbeiten der drei Forschergruppen auf Basis ihrer wissenschaftlichen Grundlage mithilfe der Handlungsregulationstheorie (siehe Hacker, 1986) und Default-Interventionist-Zwei-Prozess-Theorien (siehe Evans & Stanovich, 2013) metatheoretisch verknüpft. Zudem wurde eine Messdefinition gebildet, auf deren Basis User Interfaces (speziell 3D-CUIs) evaluiert und vorhandene Evaluationsmethoden des Feldes klassifiziert werden können. Auf diese Weise wurde die Grundlage für die Beantwortung der zentralen Frage dieser Dissertation (d.h. „Lässt sich die intuitive Benutzung von User Interfaces sowohl formativ als auch summativ mit möglichst hoher zeitlicher Anwendungseffizienz evaluieren?“) geschaffen.

Da sich die drei Hauptforschergruppen lediglich bei der *überwiegend unbewussten Anwendung von handlungsrelevantem Vorwissen* und der daraus resultierenden *zufriedenstellenden Benutzung* einer Meinung sind und bei den anderen Merkmalen lediglich immer nur maximal zwei der Forschergruppen übereinstimmen (siehe Teilabschnitt 2.1.4), wurde sich in Abschnitt 2.1 des zweiten Kapitels zunächst der genauen Spezifizierung, was unter handlungsrelevantem Vorwissen zu verstehen ist, auf Basis der Handlungsregulationstheorie und der von den Forschergruppen genutzten Literatur gewidmet (siehe Teilabschnitt 2.1.1). Durch die dadurch eingenommene Handlungsperspektive und die Auffassung von intuitiver Benutzung als Subkonzept von Usability wurde argumentiert, dass es sich bei intuitiver Benutzung als Handlung (siehe Hacker, 1986) demnach um ein zielgerichtetes Verhalten in einem bestimmten Nutzungskontext handelt. Effektivität sollte laut der IUUI- und INTUI-Forschergruppe allein schon aus pragmatischen Gründen als Merkmal zur Beurteilung von intuitiver Benutzung berücksichtigt werden, um pseudo-intuitive Handlungen nicht fälschlicherweise als intuitive Handlungen zu klassifizieren (siehe Ullrich, 2014).

Des Weiteren konnten auch die anderen objektiven (z.B. zeitliche Effizienz bei der kognitiven Informationsverarbeitung) und subjektiven Merkmale (z.B. geringe wahrgenommene Verbalisierbarkeit der kognitiven Informationsverarbeitung) von intuitiver Benutzung durch diese metatheoretische Betrachtung über die Forschergruppen und deren genutzte wissenschaftliche Grundlagen hinweg eindeutig definiert werden. Im Zuge dessen konnte ebenfalls geklärt werden, welche Merkmale geeignet sind, um das Ausmaß intuitiver Benutzung beschreiben zu können. Wie aufgrund der Tatsache, dass die drei Forschergruppen nicht einstimmig die gleichen objektiven und subjektiven Merkmale mit intuitiver Benutzung in Verbindung bringen, vermutet werden kann, ließ sich erstmalig im Forschungsbereich zu intuitiver Benutzung demonstrieren, dass sich alle von den Hauptforschergruppen vorgeschlagenen objektiven Merkmale auf Basis der Handlungsregulationstheorie und ergänzender Literatur zumindest phänomenologisch mit der mentalen Beanspruchung des Arbeitsgedächtnisses in Verbindung bringen lassen.

Die von den Forschergruppen vorgeschlagenen Merkmale stellen damit verschiedene operationale Zugänge der mentalen Beanspruchung und damit ihrer objektiven Korrelate dar. Die QUT-Forschergruppe, die zwar mentale Beanspruchung nicht direkt als Merkmal in ihrer Definition berücksichtigt, versucht beispielsweise indirekt dieses Merkmal anhand seiner objektiven Korrelate wie der zeitlichen Effizienz und der geringen Verbalisierbarkeit bei der kognitiven Informationsverarbeitung zu erfassen. Andere Forschergruppen wie die IUUI-Forschergruppe berücksichtigen stattdessen das Merkmal direkt. Mithilfe der im

Rahmen dieser Dissertation geleisteten metatheoretischen Betrachtung konnte dementsprechend eine Erklärung für die Uneinigkeit im Forschungsfeld bezüglich der objektiven Charakterisierung intuitiver Benutzung als theoretischer Forschungsbeitrag gegeben werden.

Auch bezüglich der subjektiven Merkmale, die die Zufriedenstellung bei intuitiver Benutzung erfassen, konnte auf Basis der Handlungsregulationstheorie und weiterer Literatur im Rahmen dieser Dissertation erstmalig gezeigt werden, dass sich auch alle subjektiven Merkmale mit einem bestimmten intuitive Benutzung charakterisierenden Merkmal, dem zufriedenstellenden Gefühl von Flüssigkeit, zumindest phänomenologisch in Verbindung bringen lassen und damit als operationale Zugänge bzw. subjektive Korrelate dieses Gefühls fungieren können. Die IUUI-Forschergruppe beispielsweise nähert sich diesem Gefühl über die wahrgenommene Vertrautheit bei der Systemnutzung an, wohingegen die anderen beiden Forschergruppen dieses unter anderem über die wahrgenommene Verbalisierbarkeit der kognitiven Informationsverarbeitung erfassen. Lediglich Forschergruppen unabhängige Forscher wie O'Brien et al. (2008) thematisieren dieses metakognitive Gefühl und die damit verbundene Literatur in ihrer Arbeit direkt. Mithilfe der im Rahmen dieser Dissertation geleisteten metatheoretischen Betrachtung konnte dementsprechend auch bezüglich subjektiver Merkmale eine Erklärung für die Uneinigkeit im Forschungsfeld bezüglich der subjektiven Charakterisierung intuitiver Benutzung als theoretischer Forschungsbeitrag gegeben werden.

Obwohl die Handlungsregulationstheorie von Hacker (1986) zwar helfen konnte (1) eine eindeutige Terminologie für die mit intuitiver Benutzung in Verbindung gebrachten Merkmale zu schaffen, (2) die von den Forschergruppen vorgeschlagenen Merkmale mit einem objektiven Merkmal (d.h. mentale Beanspruchung), einem subjektiven Merkmal (d.h. Gefühl von Flüssigkeit) und einem pragmatischen Merkmal (d.h. Effektivität) in Verbindung zu bringen, sowie (3) die unbewusste Anwendung von handlungsrelevantem Vorwissen als Grundvoraussetzung intuitiver Benutzung als *überwiegend unbewussten kognitiven Informationsverarbeitungsprozess auf Basis handlungsrelevanten Vorwissens an der Grenze zum menschlichen Bewusstsein* genauer zu spezifizieren, können die von dieser Theorie vorgeschlagenen Handlungsregulationsebenen (d.h. Unterscheidung zwischen intellektueller, perzeptiv-begrifflicher und sensomotorischer Regulationsebene) keine Auskunft darüber geben, mit welchem Mechanismus genau der Wechsel zwischen bewusster (d.h. Akkommodation durch Typ 2 Prozesse) und unbewusster (d.h. Assimilation durch Typ 1 Prozesse) kognitiver Informationsverarbeitung stattfindet (siehe Frese & Zapf, 1994; Hacker & Sachse, 2013; Nerdinger et al., 2014; Zacher & Frese, 2018).

Aufgrund der fehlenden Thematisierung des genauen Mechanismus kann die Handlungsregulationstheorie als Metatheorie ohne Zuhilfenahme ergänzender, von den jeweiligen Forschergruppen genannter, empirischer Befunde alleine keine theoretische Erklärung liefern, warum die im Forschungsfeld zu intuitiver Benutzung genannten Korrelate mit dem objektiven (d.h. mentale Beanspruchung), subjektiven (d.h. Gefühl von Flüssigkeit) und pragmatischen (d.h. Effektivität) Merkmal intuitiver Benutzung in Verbindung gebracht werden können (siehe Abschnitt 2.1). In diesem Zusammenhang konnte die Handlungsregulationstheorie ebenfalls nicht alleine theoretisch klären, ob sich das Ausmaß intuitiver Benutzung gleichermaßen in allen Merkmalen (wie von der IUUI- und QUT-Forschergruppe

angenommen) oder nur in bestimmten Merkmalen in deren relativer Ausprägung zueinander (wie von der INTUI-Forschergruppe angenommen) widerspiegelt. Daher wurden im folgenden Abschnitt 2.2 die Klasse von Default-Interventionist-Zwei-Prozess-Theorien (siehe Evans & Stanovich, 2013) als zweite in dieser Dissertation genutzte Metatheorie vorgestellt und die Flüssigkeit präattentiver metakognitiver Prozesse als Mechanismus für den Wechsel zwischen den beiden archetypischen Verarbeitungsmodi herausgearbeitet.

Um diese beiden offenen Fragen letztendlich durch die theoretische Analyse des Mechanismus des dynamischen Wechsels zwischen Assimilation und Akkommodation klären zu können, wurde das Drei-Instanzen-Modell des Geistes (siehe Stanovich et al., 2014) vorgestellt und erläutert, dass bei jeder Handlungsregulation zunächst immer der autonome Geist (d.h. unbewusste Assimilation durch Typ 1 Prozesse) eine initiale, präattentive Handlungstendenz abgibt. Dabei enthält diese Handlungstendenz nicht nur den Inhalt selbst, sondern auch eine präattentive metakognitive Komponente, die als Gefühl von Flüssigkeit bezeichnet wird. Diese Komponente entsteht aus einem präattentiven, affektiven Richtimpuls während der Assimilation. Anhand dieses Gefühls können präattentive, metakognitive Überwachungsprozesse, die im Modell zusammenfassend als reflektierender Geist bezeichnet werden, entscheiden, ob die vom autonomen Geist geleistete kognitive Informationsverarbeitung zur Lösung des der Handlung zugrunde liegenden Problems zufriedenstellend ist. Bei einem starken Gefühl muss der reflektierende Geist die initiale Antworttendenz des autonomen Geistes nicht durch ein kognitives Abkoppeln (d.h. bewusste Akkommodation durch Typ 2 Prozesse) überschreiben. Bei einem schwachen Gefühl muss der reflektierende Geist hingegen ein kognitives Abkoppeln durch bewusste Akkommodation initiieren, was letztendlich dann vom algorithmischen Geist ausgeführt wird (siehe Stanovich et al., 2014). Für das kognitive Abkoppeln werden die in diesem Kapitel vorgestellten exekutiven Grundfunktionen des Arbeitsgedächtnisses (d.h. Updating, Shifting, Inhibition) benötigt, weswegen sich dessen Ausmaß in der mentalen Beanspruchung des Handelnden widerspiegelt. Denn je schwächer das Gefühl von Flüssigkeit und somit der Bedarf an kognitivem Abkoppeln ist, umso mehr bewusste Akkommodationsprozesse müssen für die Lösung des Handlungsproblems aufgewendet werden, was sich gleichermaßen in einer erhöhten mentalen Beanspruchung des Handelnden erkennbar macht.

Im Anschluss wurde die Quintessenz dieser kritischen Auseinandersetzung in Form einer Messdefinition festgehalten. Sie versteht intuitive Benutzung als das Ausmaß, mit dem ein Produkt effektiv und mental effizient genutzt wird, was mit einem starken metakognitiven Gefühl von Flüssigkeit einhergeht. Im Zuge dessen wurden die von den verschiedenen Forschergruppen vorgeschlagenen Merkmale auch mithilfe von Default-Interventionist-Zwei-Prozess-Theorien als Korrelate unter dem dazugehörigen zentralen objektiven Merkmal (d.h. mentale Beanspruchung) und dem zentralen subjektiven Merkmal (d.h. Gefühl von Flüssigkeit) gruppiert. Demzufolge konnte der im Rahmen dieser Dissertation geleistete theoretische Forschungsbeitrag die drei zentralen Merkmale herausarbeiten, an denen sich zumindest theoretisch das Ausmaß intuitiver Benutzung mit hoher wissenschaftlicher Güte beurteilen lassen sollte. Damit konnte auch die erste reine Messdefinition im Forschungsfeld zu intuitiver Benutzung bereitgestellt werden. Basierend auf dieser Messdefinition konnte das objektive Charaktermerkmal (d.h. mentale Beanspruchung), das subjektive Charaktermerkmal (d.h. Gefühl von Flüssigkeit) und das pragmatische Charaktermerkmal (d.h. Effektivität) in Kapitel 3 sowohl mit formativen als auch mit summativen Evaluationsmethoden in Verbindung gebracht werden. Auf diese Weise konnte im Vergleich

zu früheren theoretischen Betrachtungen (siehe Blackler et al., 2018), die formative Evaluationsmethoden nur am Rande betrachteten und bei summativen Evaluationsmethoden lediglich zwischen objektiven und subjektiven Methoden unterschieden, eine detailliertere Klassifikation von Evaluationsmethoden durch die Verknüpfung zwischen Merkmal und Methode erreicht werden. Beispielsweise bilden objektive Methoden überwiegend die mentale Beanspruchung des Nutzers als objektives Merkmal ab. Dies kann Forscher bei der künftigen Auswahl ihrer Methoden und der Operationalisierung intuitiver Benutzung unterstützen.

Unabhängig von der empirischen Gültigkeit der Messdefinition, die in Abschnitt 7.6 zusammenfassend diskutiert wird, kann sich zukünftige Forschung drei weiteren Themen widmen. Im Gegensatz zu mentaler Beanspruchung, die im HCI-Bereich gut erforscht ist (siehe Cain, 2007; F. Chen et al., 2016; Young et al., 2015), ist die genaue Phänomenologie des Gefühls von Flüssigkeit außerhalb der HCI noch relativ unerforscht (siehe Ackerman & Thompson, 2017). Es ist an dieser Stelle weiter zu klären, welche in Teilabschnitt 2.1.4 noch nicht aufgeführten Indikatoren das Gefühl von Flüssigkeit erfassen können.

Des Weiteren wurden im Rahmen dieser Dissertation und damit auch bei der Messdefinition sog. Traits (d.h. Denkdispositionen) nicht berücksichtigt. Traits sind definiert als zeitstabile, überdauernde Merkmale und Eigenschaften, welche einen Menschen prädisponieren, sich in verschiedenen Situationen durchgehend konsistent zu verhalten (Gerrig & Zimbardo, 2008). Laut der in Abschnitt 2.2 vorgestellten Default-Interventionist-Zwei-Prozess-Theorien (siehe Evans & Stanovich, 2013; Stanovich, 1999; Stanovich et al., 2014) kann sich das Gefühl von Flüssigkeit und damit auch die mentale Beanspruchung des Handelnden unterschiedlich stark einstellen, was wiederum individuell von den intellektuellen Fähigkeiten des jeweiligen Menschen abhängig ist. Traits sollten in diesem Zusammenhang von zukünftigen theoretischen Arbeiten im Forschungsbereich zu intuitiver Benutzung thematisiert werden.

Schließlich könnte sich zukünftige Forschung nicht nur dem Gefühl von Flüssigkeit, sondern auch der mentalen Beanspruchung des Arbeitsgedächtnisses direkt widmen. Im Gegensatz zur reinen Definition von mentaler Beanspruchung, die in dieser Dissertation in Anlehnung an Young et al. (2015) als das Ausmaß an benötigten kognitiven Ressourcen des Arbeitsgedächtnisses (d.h. Kurzzeitgedächtnisses) definiert wurde, welche unter Berücksichtigung des Nutzungskontextes zur effektiven Erfüllung des Handlungsziels nötig sind, stimmen eine Reihe von Forschern in der HCI darin überein, dass heutzutage sehr viele vage Definitionen mentaler Beanspruchung existieren. Diese vagen Definitionen erschweren das Verständnis für Evaluatoren in verschiedenen Kontexten unnötig (Longo, 2014, 2015). Mentale Beanspruchung ergibt sich immer dynamisch aus der Interaktion der (1) Anforderungen der Aufgabe, (2) den Umständen, unter denen diese Aufgabe stattfindet, und den (3) Fähigkeiten, (4) dem Verhalten und der (5) Wahrnehmung des Nutzers (S. G. Hart & Staveland, 1988). Daher raten aktuelle Arbeiten, wie die von Longo (2014, 2015), mentale Beanspruchung zukünftig mithilfe eines mathematischen Modells vorherzusagen. Wickens (2017) räumt solchen Computermodellen in der HCI zukünftig eine große Wichtigkeit ein, weswegen sich auch das Forschungsfeld der intuitiven Benutzung diesen Modellen widmen sollte.

7.4 Methodische Forschungsbeiträge

Methodische Forschungsbeiträge beschäftigen sich mit Methoden der Wissenschaftler oder Praktiker in der HCI. An dieser Stelle ist anzumerken, dass gänzlich neue Methodenvorschläge selten sind und Weiterentwicklungen oder Adaptionen von bereits verfügbaren Methoden häufiger erfolgen (Wobbrock & Kientz, 2016). Methodische Forschungsbeiträge sollten immer bezüglich der Neuartigkeit und Nützlichkeit, der darin aufgeführten neuen oder verbesserten Methoden beurteilt werden und so ausführlich beschrieben sein, dass diese von Experten aus der Wissenschaft oder Praxis angewendet werden können (Wobbrock & Kientz, 2016). Der Nachweis der Neuartigkeit erfolgt üblicherweise theoretisch und der Nachweis der Nützlichkeit üblicherweise empirisch formal anhand der in Kapitel 3 genannten formativen und summativen Gütekriterien durch eine Meta-Evaluation oder nicht formal durch eine Fallstudie. Das heißt die Methode wird unabhängig von Außenkriterien in einem bestimmten Setting angewendet und die Ergebnisse beispielsweise bezüglich der in Teilabschnitt 3.2.5 beschriebenen nicht formalen Gütekriterien analysiert. Auf diese Weise können potentielle Anwender überzeugt werden, dass die getestete Methode eine ausreichende wissenschaftliche und praktische Güte besitzt, um diese im wissenschaftlichen und/oder praktischen Kontext anwenden zu können (siehe Wobbrock & Kientz, 2016).

7.4.1 IntuiBeat-S zur summativen Evaluation intuitiver Benutzung

Der erste methodische Forschungsbeitrag dieser Dissertation bestand in der *Entwicklung und der späteren empirischen Meta-Evaluation der für die summative Evaluation intuitiver Benutzung verwendeten Methode IntuiBeat-S*. Wie bereits in Abschnitt 3.6 ausführlich erläutert wurde, verfügt der im Forschungsfeld verfügbare summative Benchmark, die CHAI-Methode, als objektive Methode über die nötige wissenschaftliche Güte, um User Interfaces (speziell 3D-CUI-Interaktionslösungen im Anwenderprojekt 3D-GUIde) summativ evaluieren zu können. Der Benchmark weist jedoch aufgrund der damit verbundenen relativ zeitaufwendigen Anwendung des CHAI-Beurteilungsschemas zur retrospektiven Bewertung der Intuitivität nicht die aus Anwenderprojektsicht gewünschte, hohe zeitliche Anwendungseffizienz auf. In Folge dessen wurde auf Grundlage der in Abschnitt 2.2.2 abgeleiteten Messdefinition, die mentale Beanspruchung als charakterisierendes objektives Merkmal intuitiver Benutzung identifiziert, und auf Basis der Tatsache, dass in Abschnitt 3.6 auf Inhibition basierenden Zweitaufgabenleistungsmaßen ein hohes Potential zur Evaluation mentaler Beanspruchung zugesprochen wurde, eine solche Zweitaufgabe als potentieller summativer Benchmark für die Evaluation intuitiver Benutzung untersucht. Die ursprünglich aus der Lernforschung stammende Rhythmusmethode von Park und Brünken (2015) wurde daher als Grundlage für die im Rahmen dieser Dissertation vorgestellte Methode *IntuiBeat-S* ausgewählt.

Im Gegensatz zur Rhythmusmethode (Park & Brünken, 2015), deren wissenschaftliche Güte aufgrund problematischer Quasi-Außenkriterien (z.B. ausgewählte subjektive Maße und okulare physiologische Maße, deren wissenschaftliche Güte selbst anzuzweifeln ist) zwar teilweise, aber nicht vollständig formal auf Basis der drei wissenschaftlichen Hauptgütekriterien nachgewiesen wurde (siehe Abschnitt 4.1), konnte mit der im Rahmen

des ersten Experiments in Abschnitt 5.1 beschriebenen Vorgehensweise bei der Durchführung und Datenauswertung mithilfe von IntuiBeat-S und den damit verbundenen Setups (d.h. IntuiBeat-Software und entsprechender Hardware, siehe Abschnitt 7.5) erstmalig eine für den HCI-Bereich geeignete Adaption der Rhythmusmethode für die summative Evaluation intuitiver Benutzung als methodischer Forschungsbeitrag erbracht werden. Neben dieser Adaption und damit der Erfüllung der von Wobbrock und Kientz (2016) geforderten Neuartigkeit eines methodischen Forschungsbeitrags, lieferte die in Kapitel 5 beschriebene, auf Basis von drei Experimenten durchgeführte, Meta-Evaluation erste empirische Befunde (siehe Teilabschnitt 7.6.1) bezüglich der Güte (d.h. wissenschaftliche Güte und zeitliche Anwendungseffizienz) von inhibitionsbasierten Rhythmuszweitaufgaben (d.h. IntuiBeat-S) als summative Evaluationsmethode für intuitive Benutzung. Infolgedessen konnte IntuiBeat-S im Projekt 3D-GUIde als neuer Benchmark zur summativen Evaluation von 3D-CUI-Interaktionslösungen angewendet werden (siehe Burmester et al., 2018) und damit die von Wobbrock und Kientz (2016) für einen methodischen Forschungsbeitrag geforderte Nützlichkeit aus Anwenderprojektsicht erfüllen. Trotzdem weist die Meta-Evaluation von IntuiBeat-S Limitationen auf, deren Untersuchung Gegenstand zukünftiger Forschung sein kann. Auf diesen Aspekt wird im Rahmen der Diskussion des empirischen Forschungsbeitrags, der Meta-Evaluation von IntuiBeat-S, eingegangen (siehe Teilabschnitt 7.6.1).

7.4.2 IntuiBeat-F zur formativen Evaluation intuitiver Benutzung

Der zweite methodische Forschungsbeitrag dieser Dissertation bestand in der *Entwicklung und der späteren empirischen Meta-Evaluation der für die formative Evaluation intuitiver Benutzung genutzten Methode IntuiBeat-F*. Wie bereits in Abschnitt 3.2.3 ausführlich erläutert, verfügt der im Forschungsfeld verfügbare formative Benchmark, der Nutzertest mit retrospektivem Think-Aloud-Protokoll, zwar über die nötige wissenschaftliche Güte, um User Interfaces (speziell 3D-CUI-Interaktionslösungen im Anwenderprojekt 3D-GUIde) formativ evaluieren zu können, weist aber aufgrund der fehlenden Unterstützung bei der Durchführung des retrospektiven Interviews nicht die aus Anwenderprojektsicht gewünschte hohe zeitliche Anwendungseffizienz auf. Wie bereits im letzten Teilabschnitt erläutert, wurde, aufgrund des Potentials von auf Inhibition basierenden Zweitaufgaben, die Rhythmusmethode von Park und Brünken (2015) als Grundlage für die im Rahmen dieser Dissertation vorgestellte Methode *IntuiBeat-S* ausgewählt. Zwar wurde die Rhythmusmethode noch nicht zur formativen Evaluation eingesetzt, jedoch weisen ähnliche Arbeiten wie beispielsweise von M. J. Albers (2011) darauf hin, dass Klopf-Zweitaufgaben potentiell zur formativen Evaluation im Usability-Bereich gut geeignet sind. Zusammen mit der Rhythmusmethode bilden diese Arbeiten die Grundlage der im Rahmen dieser Dissertation vorgestellten Methode *IntuiBeat-F*.

Im Gegensatz zur Rhythmusmethode (Park & Brünken, 2015), deren wissenschaftliche Güte als formative Evaluationsmethode weder formal (siehe Abschnitt 3.3) noch nicht formal (siehe Abschnitt 3.2.5) untersucht wurde, und der Methode von M. J. Albers (2011), bei der ebenfalls derartige Belege fehlen (siehe Abschnitt 4.2), konnte mit der im Rahmen des vierten Experiments in Abschnitt 6.1 und des sechsten Experiments in Abschnitt 6.2 (inkl. zusätzlicher Analysesoftware) beschriebenen Vorgehensweise bei der

Durchführung und Datenauswertung mithilfe von IntuiBeat-F und den damit verbundenen Setup (d.h. IntuiBeat-Software und entsprechender Hardware, siehe Abschnitt 7.5) erstmalig eine für den HCI-Bereich geeignete Adaption der Rhythmusmethode für die formative Evaluation intuitiver Benutzung als methodischer Forschungsbeitrag erbracht werden. Neben dieser Adaption und damit der Erfüllung der von Wobbrock und Kientz (2016) geforderten Neuartigkeit eines methodischen Forschungsbeitrags lieferte die in Kapitel 6 beschriebene, auf Basis von vier Experimenten durchgeführte Meta-Evaluation erste empirische Befunde (siehe Teilabschnitt 7.6.2) bezüglich der Güte (d.h. wissenschaftliche Güte und zeitliche Anwendungseffizienz) von inhibitionsbasierten Rhythmuszweitaufgaben (d.h. IntuiBeat-F) als formative Evaluationsmethode für intuitive Benutzung. Infolgedessen konnte IntuiBeat-F im Projekt 3D-GUIde als neuer formativer Benchmark zur formativen Evaluation von 3D-CUI-Interaktionslösungen angewendet werden (siehe Burmester et al., 2018) und konnte damit auch die von Wobbrock und Kientz (2016) für einen methodischen Forschungsbeitrag geforderte Nützlichkeit aus Anwenderprojektsicht erfüllen. Jedoch weist die Meta-Evaluation von IntuiBeat-F Limitationen auf, deren Untersuchung Gegenstand zukünftiger Forschung sein kann. Auf diesen Aspekt wird im Rahmen der Diskussion des empirischen Forschungsbeitrags, der Meta-Evaluation von IntuiBeat-F, eingegangen (siehe Teilabschnitt 7.6.2).

7.5 Forschungsbeitrag im Sinne eines Werkzeugs

Unter Werkzeugen (auch als Artefakte bezeichnet) werden in der HCI jegliche Art von Erfindungen verstanden, wie beispielsweise neue Systeme, Architekturen, Techniken, Geräte oder Designs, die neue Möglichkeiten aufzeigen, neue Wirkungsweisen ermöglichen, neue Erkenntnisse und Entdeckungen fördern oder uns zwingen über mögliche Zukunftsszenarien nachzudenken (Wobbrock & Kientz, 2016). Forschungsbeiträge im Sinne eines Werkzeugs sind per Definition von bisher nicht vorhandenen Erfindungen, die in Form von Prototypen, Sketches, Mockups oder anderen ähnlichen Darstellungen vorliegen, die zumindest bis zu einem gewissen Grad die Funktionalität des Werkzeugs demonstrieren können. Derartige Forschungsbeiträge werden üblicherweise empirisch beurteilt, was aber nicht immer vorkommt und in der HCI-Praxis oft von der Art des vorgestellten Werkzeugs abhängig ist.

Neben der im Rahmen dieser Dissertation für die formative und summative Evaluation erarbeiteten Methodik wurden für die Anwendung beider Methoden mit der IntuiBeat-Software (d.h. Aufzeichnungs- und Analysesoftware) bzw. Hardware (d.h. USB-Fußpedal) auch Werkzeuge bereitgestellt. Hierbei konnte empirisch demonstriert werden, dass sowohl ein herkömmliches USB-Fußpedal (siehe Abschnitt 6.2, 6.3, 6.4, 6.1, 5.2 und 5.1), mit dem die Eingabe des Rhythmus über den Fußballen erfolgt, und ein durch additive Fertigung (d.h. 3D-Druck) hergestelltes USB-Fußpedal „Taktschuh“ (siehe Abschnitt 5.3), mit dem die Eingabe des Rhythmus mit der Ferse erfolgte, gleichermaßen zur Evaluation intuitiver Benutzung eingesetzt werden kann. Dies stellt bezogen auf den praktischen Einsatz beider Methoden einen Vorteil dar, da für die Evaluation nicht ein bestimmtes Fabrikat genutzt werden muss und der Evaluator bei der Auswahl seiner Hardware somit relativ unabhängig ist. Im Gegensatz zu den Ursprungsarbeiten zur Rhythmusmethode (siehe Korbach et al.,

2018; Park & Brünken, 2015), für die in den Experimenten verschiedene, mitunter kostspielige Setups (d.h. Audacity-Software mit einem speziell entwickelten USB-Fußpedal oder E-Prime-Software mit dazugehörigem USB-Fußpedal) zur Rhythmusaufzeichnung genutzt wurden (siehe Abschnitt 5.1), kann der Evaluator bei dem im Rahmen dieser Dissertation bereitgestellten Werkzeug entweder auf gewöhnliche USB-Fußpedale zurückgreifen oder sich das am Lehrstuhl für Psychologische Ergonomie entwickelte USB-Fußpedal „Taktschuh“ selbst mit einem 3D-Drucker herstellen (siehe Abschnitt 5.3).

Zusätzlich wurde mit der IntuiBeat-Software ein standardisiertes Tool zur Aufzeichnung und Auswertung von Rhythmusabweichungen bereitgestellt. Der Evaluator wird hierbei im Gegensatz zu früheren Setups (siehe Korbach et al., 2018; Park & Brünken, 2015) aktiv (d.h. durch entsprechende Labels, Buttons und Timer) bei der Erhebung der Baseline des Nutzers (durch den Baseline-Modus der Software) und bei der Durchführung des eigentlichen Nutzertests (durch den Experimental-Modus der Software) unterstützt. Dies war bei früheren Setups nicht möglich, da für die Aufzeichnung Softwarepakete eingesetzt wurden (siehe Abschnitt 5.1), die nicht speziell für die Aufzeichnung von Rhythmusabweichungen entwickelt wurden (z.B. Audacity ist eine allgemeine Audioaufzeichnungssoftware). Des Weiteren wird bei der IntuiBeat-Software eine geringere Zeit für die Einarbeitung für den Evaluator benötigt (d.h. Schulungsaufwand ist geringer aufgrund besserer Usability durch eine bessere Passung von Aufgabe, System und Nutzer). Auch im Vergleich mit anderen im Forschungsfeld zu intuitiver Benutzung verfügbaren objektiven Methoden (siehe Teilabschnitt 3.6.2), die alle nicht über eine standardisierte Möglichkeit der Erhebung (z.B. es existieren verschiedene Softwarepakete, um Fixationen mit einem Eye-Tracker aufzuzeichnen) und der Auswertung (z.B. die CHAI-Methode stellt keine Softwareunterstützung für die Beurteilung der Intuitivität von Klicks bereit, sondern lediglich ein Bewertungsschema) verfügen, kann das im Rahmen dieser Dissertation bereitgestellte Werkzeug den Evaluator gezielt unterstützen.

Daneben stellt die IntuiBeat-Software durch die zusätzlich für die retrospektive Befragung bei der formativen Evaluation entwickelte Analysesoftware an den richtigen Stellen der Videoaufzeichnung die mithilfe des AMPD-Algorithmus (siehe Scholkmann et al., 2012) identifizierten, kritischen Ereignisse bereit, die der Evaluator während des Interviews als Anhaltspunkte bei der Ableitung von Nutzungsproblemen verwenden kann (siehe Abschnitt 6.3). Für die summative Evaluation stellt die IntuiBeat-Software entsprechend, die mittlere Rhythmusabweichung zur Verfügung (d.h. Abweichung des langen und kurzen Rhythmusintervalls in Abhängigkeit von der zuvor ermittelten Baseline des Nutzers). Darüber hinaus ermöglicht sie den Zugang zu den dieser Metrik zugrunde liegenden Daten in Form von verschiedenen Dateien (z.B. Base-Datei für Informationen zur Baseline-Messung). Die von der im Rahmen dieser Dissertation als Werkzeug bereitgestellten IntuiBeat-Software und IntuiBeat-Hardware geleistete, Standardisierung ist dementsprechend insgesamt im Vergleich zu früheren Setups (siehe Korbach et al., 2018; Park & Brünken, 2015) und anderen Evaluationmethoden für intuitive Benutzung (siehe Kapitel 3) hoch. Daher sollte dieses neue Setup zur Erfassung und Auswertung von Rhythmusabweichungen zumindest theoretisch die Objektivität der Evaluation intuitiver Benutzung auf Basis von Rhythmusabweichungen verbessern.

Trotzdem ergeben sich eine Reihe von Verbesserungsmöglichkeiten, denen sich Forschung zukünftig widmen kann. Zunächst ist anzumerken, dass, egal ob für die Aufzeichnung

der Rhythmusabweichungen ein herkömmliches Fußpedal oder der gedruckte „Taktschuh“ genutzt wird, der Evaluator immer darauf angewiesen ist, dass die vom Probanden getragenen Schuhe eine Aufzeichnung zulassen. Der Rhythmus kann beispielsweise mit keinem der beiden Fußpedale zuverlässig aufgezeichnet werden, wenn Probanden Schuhe mit hohen Absätzen tragen. Des Weiteren kann insbesondere die Schuhgröße des Probanden (d.h. zu kleine oder zu große Füße) speziell beim „Taktschuh“, der um die Ferse geschnallt werden muss (siehe Abschnitt 5.3), die Aufzeichnungsqualität negativ beeinflussen. Um diese Problematik zu minimieren, wurden die Versuchsteilnehmer in allen Experimenten darum gebeten für die Erhebung flache Sportschuhe zu tragen. Künftige Forschung könnte diese Problematik umgehen, indem statt eines Drucksensors auf dem Fußpedal ein Bewegungssensor direkt auf dem Fuß genutzt wird, durch den festgestellt werden kann, inwiefern der Proband seinen Fußballen oder seine Ferse zur Rhythmuseingabe angehoben und abgesenkt hat.

Aktuelle Arbeiten beschäftigen sich insbesondere im medizinischen Kontext zur neurologischen Überprüfung der menschlichen Motorik und damit der Integrität der Komponenten des zentralen Nervensystems mit einem derartigen Setup. Eine aktuelle Übersicht über die dabei verwendete Hardware und damit verbundenen Schnittstellen für Softwareanwendungen geben Đurić-Jovičić et al. (2018). Im Rahmen dieser Übersicht wird auch die Erfassung von Rhythmusabweichungen mithilfe von anderen Extremitäten (z.B. Finger) und die dafür erforderlichen technischen Komponenten beschrieben. Da die mithilfe eines gewöhnlichen USB-Fußpedals oder des „Taktschuhs“ vorgenommene Aufzeichnung eigentlich nur im Sitzen zuverlässig funktionieren kann und sich damit nicht für die Evaluation von mobilen Geräten (z.B. Evaluation von Smartphones) oder für die Evaluation in virtuellen Umgebungen (z.B. Evaluation für VR-Headsets) eignet, sollte sich zukünftige Forschung schließlich auch der Exploration und der anschließenden empirische Meta-Evaluation von IntuiBeat-F bzw. IntuiBeat-S unter diesen Anforderungen widmen.

Neben der im Rahmen dieser Dissertation genutzten Hardware kann auch die Intui-Beat-Software weiterentwickelt werden. Die für die formative Evaluation benötigte Analysesoftware kann beispielsweise auch direkt in die Aufzeichnungssoftware integriert werden, sodass der Evaluator nur eine Applikation verwenden muss. An dieser Stelle kann auch die manuelle Zuordnung von Videodateien zu Marker-Dateien automatisiert werden (siehe Abschnitt 6.3). Des Weiteren kann die IntuiBeat-Software durch eine Reihe von weiteren Komfortfunktionen erweitert werden. Aktuell lässt sich mit der Software nur eine Art von Rhythmus (d.h. kurzes Rhythmusintervall von 500 Millisekunden und langes Rhythmusintervall von 1500 Millisekunden) erfassen. Daneben kann die Reaktion auf den Ausgabewert des USB-Fußpedals (z.B. Ausgabe einer „0“ auf der Tastatur, was dann von der IntuiBeat-Software weiterverarbeitet wird) nur programmatisch geändert werden. Zudem stellt der Speicherort für die ausgegebenen Dateien (z.B. Raw-Datei und Clean-Datei) immer den Ordner dar, in dem sich auch die JAR-Datei der IntuiBeat-Aufzeichnungs- bzw. Analysesoftware befindet. Diese und andere Parameter sollten sich in einer späteren Version direkt über die IntuiBeat-Software ändern lassen. Darüber hinaus wird dem Evaluator nach der Messung nur eine Metrik (die mittlere Rhythmusabweichung des kurzen und langen Rhythmusintervalls) angezeigt, was zukünftig noch durch weitere Metriken erweitert werden kann (z.B. Anzahl gültiger Rhythmusabweichungen, mittlere Abweichungen von Rhythmus-Peaks, die bei der formativen Evaluation die kritischen Ereignisse markieren). Da im Rahmen dieser Dissertation bei der Evaluation verhindert werden sollte, dass

es zu ungewollten wechselseitigen Beeinflussungen der Bildschirmaufzeichnungssoftware und der Rhythmusaufzeichnungssoftware kommt (d.h. etwaige Wettlaufsituationen bei der Nutzung von Ressourcen wie Prozessorkernen, siehe Netzer & Miller, 1992), wurden zwei getrennte PCs für die Aufzeichnung genutzt. Zukünftige Forschung könnte untersuchen, wie künftig die Aufzeichnung vollständig auf dem Gerät des Nutzers laufen kann und der Evaluator dementsprechend weniger Hardware für den Einsatz der Methode benötigt.

7.6 Empirische Forschungsbeiträge

Empirische Forschungsbeiträge beschreiben neue Befunde, die auf systematisch beobachteten Daten basieren (Wobbrock & Kientz, 2016). Solche Forschungsbeiträge können entweder quantitativ oder qualitativ (oder gemischt) sein, und sind üblicherweise das Ergebnis wissenschaftlicher Studien unterschiedlicher Art (z.B. Laborstudien, Feldstudien, ethnographische Studien). In der HCI ermöglichen empirische Forschungsbeiträge auf Basis neuer Befunde neue Erkenntnisse über das menschliche Verhalten und dessen Zusammenhang mit Technologie (Wobbrock & Kientz, 2016). In der HCI werden als empirische Forschungsmethoden laut Wobbrock und Kientz (2016) unter anderem Laborexperimente, Feldstudien, Interviews, Fokusgruppen, Umfragen, Nutzertests (z.B. in Form eines Usabilitytests), Fallstudien, Tagebuchstudien, Nutzungskontextanalysen und automatisches Sammeln verschiedener Daten (z.B. automatisches Logging von Ereignissen) verstanden. Empirische Forschungsbeiträge werden üblicherweise als vertrauenswürdig eingestuft, wenn die für die Erhebung genutzten Methoden über eine ausreichende Güte verfügen und sie gewissenhaft und präzise angewendet wurden. Laut Wobbrock und Kientz (2016) werden empirische Forschungsbeiträge immer dann positiv aufgenommen, wenn die verwendeten Methoden fundiert sind und es sich für das Forschungsfeld der HCI um wichtige empirische Ergebnisse handelt.

7.6.1 Meta-Evaluation von IntuiBeat-S

Die erste Forschungsfrage dieser Dissertation widmete sich der wissenschaftlichen Güte von IntuiBeat-S als summative Evaluationsmethode für intuitive Benutzung auf Basis der drei formalen Hauptgütekriterien *Objektivität*, *Reliabilität* und *Validität*. Mithilfe der ersten drei Experimente (d.h. Experimente 1 - 3) wurde bei verschiedenen Paaren von CUIs (d.h. weniger intuitiv benutzbares vs. stärker intuitiv benutzbares CUI) die wissenschaftliche Güte von IntuiBeat-S im Vergleich mit den in Abschnitt 3.6 identifizierten Außenkriterien (d.h. Quasi-Außenkriterien und der aktuell im Forschungsfeld verfügbare summative Benchmark) formal mit insgesamt 78 Versuchspersonen gezeigt. Das Hauptgütekriterium der Objektivität wurde in allen drei Experimenten nicht empirisch nachgewiesen, da laut Moosbrugger und Kelava (2007), sowie Döring und Bortz (2016) einer Methode *Objektivität* in Form von Durchführungs, Auswertungs- und Interpretationsobjektivität attestiert werden kann, wenn eine Methode nahezu standardisiert durchgeführt wird. Im Rahmen dieser Dissertation wurde angenommen, dass diese Anforderung durch den Einsatz der IntuiBeat-Software erfüllt wurde, da diese die Durchführung und Auswertung von IntuiBeat-S im Bereich summativer Evaluation nahezu standardisiert gestattet

(d.h. Berücksichtigung der genauen Arbeitsschritte durch die bereitgestellten Funktionalitäten der Software und Bereitstellung von standardisierten Auswertungsdateien) und bei der Interpretation der Ausgabe kein besonderes Expertenwissen voraussetzt (d.h. niedrige intuitive Benutzung bei hohen Rhythmusabweichungen von der Baseline, hohe intuitive Benutzung bei niedrigen Rhythmusabweichungen von der Baseline).

Beim ersten Experiment wurden für den empirischen Nachweis der Hauptgütekriterien *Reliabilität* und *Validität* mithilfe von IntuiBeat-S und verschiedenen Quasi-Außenkriterien eine weniger intuitiv benutzbare Software und eine stärker intuitiv benutzbare Software miteinander verglichen (siehe Abschnitt 5.1). Die Ergebnisse zeigten, dass IntuiBeat-S das wissenschaftliche Hauptgütekriterium der Reliabilität und die divergente Validität als Teilaspekt des wissenschaftlichen Hauptgütekriteriums der Validität attestiert werden konnte. Bei der Überprüfung der konvergenten Validität konnte jedoch bei kleiner Effektstärke kein signifikanter Zusammenhang mit der Effektivität festgestellt werden, weswegen IntuiBeat-S zu diesem Zeitpunkt noch keine konvergente Validität attestiert werden konnte. Es wurde zusätzlich aus diesem Ergebnis geschlossen, dass womöglich die im ersten Experiment genutzte Operationalisierung der Effektivität zu grob war, da sie nur die erfolgreich abgeschlossenen Aufgaben berücksichtigte. Des Weiteren wurde vermutet, dass es generell zu einer Konfundierung durch den als Datengrundlage genutzten Nutzertest mit paralleler Rhythmuszweitaufgabe gekommen sein kann. Da alle konvergenten Quasi-Außenkriterien auf Basis der Daten eines Nutzertests mit zusätzlicher paralleler Rhythmusaufgabe erhoben wurden (d.h. die Voraussetzung, um IntuiBeat-S einsetzen zu können), könnte es auch zu Konfundierungen dieser Maße aufgrund der Rhythmuszweitaufgabe und einer damit verbundenen Intrusion gekommen sein. Alle im Rahmen des ersten Experiments ermittelten Effekt- und Teststärken lagen jedoch überwiegend im oberen Bereich. Des Weiteren konnten Konfundierungen durch unterschiedliche Vorerfahrung bei der Nutzung von CUIs und durch unterschiedliche rhythmische Wahrnehmung mit hoher Wahrscheinlichkeit ausgeschlossen werden.

Um zu überprüfen, ob eine derartige Konfundierung vorliegt und eine feinere Operationalisierung der Effektivität einen positiven Einfluss auf die konvergente Validität hat, wurde das zweite Experiment als konzeptuelle Replikation des ersten durchgeführt. Dabei wurden als neuer Faktor zwei andere unterschiedlich intuitiv benutzbare CUIs unter Berücksichtigung der Art des als Datengrundlage genutzten Nutzertests (d.h. Nutzertest ohne Rhythmusaufgabe vs. Nutzertest mit Rhythmusaufgabe) eingeführt und eine detailliertere Einschätzung der Effektivität bei der empirischen Überprüfung des Hauptgütekriteriums der Validität berücksichtigt (siehe Abschnitt 5.2). Die Ergebnisse demonstrierten, dass nun neben der divergenten auch die konvergente Validität (d.h. auch im Vergleich mit der Effektivität als Quasi-Außenkriterium) von IntuiBeat-S als summative Evaluationsmethode für intuitive Benutzung bestätigt werden konnte. Dies war unabhängig davon, ob als Datengrundlage ein Nutzertest mit oder ohne Rhythmusaufgabe verwendet wurde. Auf Basis dieser Ergebnisse wurde geschlussfolgert, dass das zweite Experiment die Ergebnisse des ersten Experiments konzeptuell repliziert und eine Intrusion durch die Art des als Datengrundlage verwendeten Nutzertests ausgeschlossen werden kann. Das heißt es spielt keine Rolle, ob als Datengrundlage ein Nutzertest mit parallelem Think-Aloud-Protokoll, was in Verbindung mit IntuiBeat-S einer Drittaufgabe entspricht, oder ein Nutzertest mit retrospektivem Think-Aloud-Protokoll eingesetzt wird. Konfundierungen durch Unterschiede

in der Vorerfahrung bei der Nutzung von CUIs können mit hoher Wahrscheinlichkeit ausgeschlossen werden und es lagen ferner durchgehend hohe Effektstärken vor.

Das zweite Experiment widmete sich außerdem dem die summative Evaluation betreffenden Teilaspekt der dritten Forschungsfrage dieser Dissertation, nämlich der zeitlichen Anwendungseffizienz von IntuiBeat-S im Vergleich zu bereits vorhandenen summativen Evaluationsmethoden für intuitive Benutzung. Das Ergebnis des zweiten Experiments demonstrierte hier mit ebenfalls hoher Effektstärke, dass IntuiBeat-S eine höhere zeitliche Anwendungseffizienz als der aktuelle summative Benchmark, die CHAI-Methode, besitzt. IntuiBeat-S wurde daher eine hohe zeitliche Anwendungseffizienz als wichtiger Teilaspekt praktischer Güte attestiert. Der summative Aspekt der dritten Forschungsfrage wurde dementsprechend als beantwortet befunden. In den ersten beiden Experimenten konnte damit sowohl die wissenschaftliche Güte als auch die zeitliche Anwendungseffizienz von IntuiBeat-S als summative Evaluationsmethode demonstriert werden. Jedoch wurden in beiden Experimenten Analogstichproben erhoben, die keine Kernanwender von CUIs umfassten (d.h. Studierende der Mensch-Computer-Systeme und Medienkommunikation mit geringen Vorerfahrungen bezüglich CUIs). Zudem wurde in beiden Experimenten ein USB-Fußpedal eines bestimmten Herstellers eingesetzt. Daher wurde in einem dritten Experiment überprüft, inwiefern die Ergebnisse der ersten beiden Experimente auch anhand einer anderen Stichprobe und anderer Hardware generalisiert werden können (siehe Abschnitt 5.3). Auf diese Weise konnte gewährleistet werden, dass IntuiBeat-S auch eine hohe wissenschaftliche Güte bei der Analyse von Expertennutzern (d.h. Studierende aus anderen Fächern mit mehr Vorerfahrung bei der Nutzung von CUIs) und bei Verwendung eines anderen USB-Fußpedals (d.h. Verwendung des „Taktschuhs“) attestiert werden kann.

Die Ergebnisse des dritten Experiments konnten die wissenschaftliche Güte von IntuiBeat-S als summative Evaluationsmethode für intuitive Benutzung hinsichtlich der Hauptgütekriterien *Reliabilität* und *Validität* auch unter Berücksichtigung einer im Vergleich zu den vorherigen Experimenten heterogeneren Stichprobe und eines anderen USB-Fußpedals demonstrieren. Konfundierungen durch Unterschiede in der Vorerfahrung bei der Nutzung von CUIs konnten mit hoher Wahrscheinlichkeit ausgeschlossen werden. Die Effekt- und Teststärken lagen außerdem überwiegend im oberen Bereich. Jedoch konnten bezüglich der Fragebögen NASA-RTLX und QUESI sowie des objektiven Leistungsmaßes *Effektivität* keine signifikanten Unterschiede zwischen den beiden Ausprägungen der unabhängigen Variable festgestellt werden. In diesem Zusammenhang wurde diskutiert, dass der geringe Stichprobenumfang ($N = 10$) mit hoher Wahrscheinlichkeit dafür verantwortlich war und sich deswegen die Methoden dennoch als Quasi-Außenkriterien eignen können. Diese Vermutung wurde mit einer Analyse der Teststärke untermauert. Abschließend wurde unter Betrachtung aller drei Experimente abgeleitet, dass IntuiBeat-S als summative Evaluationmethode für intuitive Benutzung wissenschaftliche Güte und zeitliche Anwendungseffizienz im 3D-CUI-Bereich attestiert werden kann.

Auf Basis der ersten drei Experimente, sich primär der Meta-Evaluation von IntuiBeat-S widmeten, wurde letztendlich die wissenschaftliche Güte und die zeitliche Anwendungseffizienz von IntuiBeat-S bei der summativen Evaluation von User Interfaces (speziell 3D-CUIs) demonstriert. Darüber hinaus konnte IntuiBeat-S als neuer summativer Benchmark des Feldes identifiziert werden. Dennoch weist der im Rahmen dieser Dissertation geleistete empirische Forschungsbeitrag, die *Meta-Evaluation von IntuiBeat-S*, eine Reihe

von Limitationen auf, denen sich zukünftige Forschung widmen sollte. Die wichtigsten der in Abschnitt 5.4 genannten und im Detail besprochenen Forschungsanregungen sind im Folgenden abschließend kurz zusammengefasst:

- Meta-Evaluation von IntuiBeat-S mit empirischem Nachweis des wissenschaftlichen Hauptgütekriteriums der Objektivität durch Maße der Beurteilerübereinstimmung (siehe Moosbrugger & Kelava, 2007).
- Meta-Evaluation von IntuiBeat-S unter Berücksichtigung weiterer Nebengütekriterien (z.B. Normierung, Zumutbarkeit, Nützlichkeit) zur Erfassung weiterer Aspekte praktischer Güte (siehe Döring & Bortz, 2016).
- Meta-Evaluation von IntuiBeat-S unter Berücksichtigung weiterer Untersuchungsgegenstände neben CUIs (z.B. Medizinprodukte), um eine Generalisierung der Ergebnisse auf weitere Untersuchungsgegenstände zu erlauben und der in der HCI vorherrschenden Übergeneralisierung von Methoden vorzubeugen (siehe Teilabschnitt 3.2.5).
- Meta-Evaluation von IntuiBeat-S mit Einsatz von Experten (d.h. Praktiker mit Berufserfahrung) anstelle von Novizen (d.h. Studierende mit Vorerfahrung bei der Nutzung von CUIs) als Anwender von IntuiBeat-S zur Generalisierung der Ergebnisse auf Evaluatoren mit höherer Vorerfahrung bei der Ausführung von Nutzertests und Domänenkenntnis (z.B. Vorerfahrung mit getesteten/ähnlichen CUIs).
- Meta-Evaluation von IntuiBeat-S mit Einsatz von Experten (z.B. Praktikern mit Vorerfahrung in bestimmten Domänen wie getesteten/ähnlichen CUIs oder getesteten/ähnlichen Medizinprodukten) anstelle von Novizen (d.h. Studierende mit Vorerfahrung in bestimmter Domäne) als mit der Methode getestete Nutzergruppe zur Generalisierung der Ergebnisse auf andere Nutzergruppen.
- Meta-Evaluation von IntuiBeat-S unter Berücksichtigung längerer Aufgaben, um zu untersuchen, inwiefern Ermüdungserscheinungen der Muskeln bei der Rhythmuseingabe die Reliabilität der Methode beeinflussen.

7.6.2 Meta-Evaluation von IntuiBeat-F

Die zweite Forschungsfrage dieser Dissertation widmete sich der wissenschaftlichen Güte von IntuiBeat-F als formative Evaluationsmethode für intuitive Benutzung auf Basis der drei formalen Hauptgütekriterien *Gründlichkeit*, *Gültigkeit* und *Zuverlässigkeit*. Mithilfe der letzten vier Experimente (d.h. Experimente 4 - 7) wurde bei verschiedenen CUIs und Webseiten, die sich bezüglich ihrer intuitiven Gestaltung unterschieden (d.h. entweder weniger intuitiv benutzbare oder stärker intuitiv benutzbar sind), die wissenschaftliche Güte von IntuiBeat-F im Vergleich mit dem in Abschnitt 3.5 identifizierten, aktuellen formativen Benchmark des Forschungsfeldes und alleinigem Quasi-Außenkriterium formal mit insgesamt 96 Versuchspersonen gezeigt. Es konnten im Zuge der Meta-Evaluation keine weiteren Außenkriterien berücksichtigt werden, da auf Basis des aktuellen Forschungsstands noch keine dedizierte formative Evaluationsmethode im Forschungsfeld zu intuitiver Benutzung existiert, die nicht ursprünglich aus der übergeordneten Usabilityforschung stammt (siehe Blackler et al., 2018). Das Hauptgütekriterium der Zuverlässigkeit wurde in allen vier Experimenten nicht empirisch untersucht, da einschlägige Forschungsliteratur (z.B. Hartson

et al., 2001; Koutsabasis et al., 2007; Makri et al., 2011) rät, sich der Zuverlässigkeit erst dann zu widmen, nachdem Gründlichkeit und Gültigkeit der Methode ausgiebig sichergestellt werden konnten. Des Weiteren wurde sich im Rahmen der Meta-Evaluation von IntuiBeat-F in allen vier Experimenten auch dem die formative Evaluation betreffenden Teilaspekt der dritten Forschungsfrage gewidmet, nämlich dem Ziel, herauszufinden, wie hoch die zeitliche Anwendungseffizienz von IntuiBeat-F im Vergleich zu bereits vorhandenen formativen Evaluationsmethoden für intuitive Benutzung ausfällt.

Beim vierten Experiment kam für die Meta-Evaluation von IntuiBeat-F eine weniger intuitiv benutzbare Software (siehe Abschnitt 6.1) und beim fünften Experiment eine stärker intuitiv benutzbare Software (siehe Abschnitt 6.2) zum Einsatz, die jeweils sowohl mit IntuiBeat-F als auch mit dem als formativer Benchmark fungierenden Nutzertest mit retrospektivem Think-Aloud-Protokoll formativ evaluiert wurden. Um zusätzlich feststellen zu können, ob sich die wissenschaftliche Güte und die zeitliche Anwendungseffizienz als für das Projekt 3D-GUIde wichtiger Teilaspekt praktischer Güte von IntuiBeat-F verändert, wenn die Methode strikt oder weniger strikt verfolgt wird, wurde dieser Faktor bei der Meta-Evaluation von IntuiBeat-F in beiden Experimenten entsprechend berücksichtigt. Bei der strikten Anwendung von IntuiBeat-F konzentriert sich der Evaluator ausschließlich auf die Rhythmus-Peaks und es werden somit nur kritische Ereignisse für die Ableitung von Nutzungsproblemen berücksichtigt, die mithilfe eines Rhythmus-Peaks entdeckt werden können. Der Evaluator nutzt hingegen bei der weniger strikten Anwendung der Methode die Rhythmus-Peaks lediglich als unverbindliche Unterstützung und es werden daher auch kritische Ereignisse für die Ableitung von Nutzungsproblemen berücksichtigt, die nicht explizit mithilfe eines Rhythmus-Peaks entdeckt werden konnten. Es konnten in beiden Experimenten jedoch in beiden Fällen (d.h. strikte und weniger strikte Anwendung von IntuiBeat-F) nur das wissenschaftliche Gütekriterium der Gültigkeit und die zeitliche Anwendungseffizienz bestätigt werden. Dabei konnten Konfundierungen durch unterschiedliche Vorerfahrung bei der Nutzung von CUIs mit hoher Wahrscheinlichkeit ausgeschlossen werden. Die Effekt- und Teststärken lagen in beiden Experimenten überwiegend im oberen Bereich.

In beiden Experimenten wurden durch Rhythmus-Peaks als kritisch eingestufte Ereignisse lediglich in Form einer Tabelle in einem Tabellenkalkulationsprogramm präsentiert. Dadurch könnte der Evaluator mit hoher hoher Wahrscheinlichkeit einige Rhythmus-Peaks bei der Zuordnung zu den entsprechenden Videostellen übersehen haben, was sich negativ auf die wissenschaftliche Güte von IntuiBeat-F ausgewirkt haben könnte. Daher wurde vorgeschlagen, dass die Zuordnung von Rhythmus-Peaks zu den dazugehörigen Videostellen mithilfe einer zusätzlichen Analysesoftware verbessert werden könnte. Somit wurde für das sechste und siebte die zusätzliche Analysesoftware in Kombination mit der IntuiBeat-Software für das retrospektive Interview eingesetzt.

Unter Berücksichtigung dieser neuen Analysesoftware wurden im sechsten Experiment für die Meta-Evaluation von IntuiBeat-F eine weniger intuitiv benutzbare Software (siehe Abschnitt 6.3) und beim siebten Experiment eine stärker intuitiv benutzbare Software (siehe Abschnitt 6.4) als Untersuchungsgegenstand eingesetzt. In beiden Fällen (d.h. strikt und weniger strikte Anwendung von IntuiBeat-F) wurden diese, wie bei den beiden Experimenten zuvor, sowohl mit IntuiBeat-F als auch mit dem als formativer Benchmark fungierenden Nutzertest mit retrospektivem Think-Aloud-Protokoll formativ evaluiert. In

beiden Fällen konnte die wissenschaftliche Güte von IntuiBeat-F nun auch hinsichtlich des wissenschaftlichen Hauptgütekriteriums der Gründlichkeit demonstriert werden. Das wissenschaftliche Hauptgütekriterium der Gültigkeit und die zeitliche Anwendungseffizienz von IntuiBeat-F konnten im Vergleich zum aktuellen formativen Benchmark auch unter Berücksichtigung der neuen Analysesoftware in beiden Fällen (d.h. strikte und weniger strikte Anwendung von IntuiBeat-F) nachgewiesen werden. In beiden Experimenten konnten Konfundierungen durch Unterschiede in der Vorerfahrung bei der Nutzung von CUIs mit hoher Wahrscheinlichkeit ausgeschlossen werden. Die Effektstärken lagen in beiden Experimenten überwiegend im oberen Bereich. Abschließend wurde unter Betrachtung aller vier Experimente abgeleitet, dass IntuiBeat-F als formative Evaluationmethode für intuitive Benutzung wissenschaftliche Güte und zeitliche Anwendungseffizienz im 3D-CUI-Bereich attestiert werden kann.

Auf Basis der letzten vier, sich primär der Meta-Evaluation von IntuiBeat-F widmeten, Experimente wurde demzufolge die wissenschaftliche Güte und die zeitliche Anwendungseffizienz von IntuiBeat-F bei der formativen Evaluation von User Interfaces (speziell CUIs und Webseiten) demonstriert. Darüber hinaus konnte IntuiBeat-F als neuer formativer Benchmark des Feldes identifiziert werden. Trotzdem weist der im Rahmen dieser Dissertation geleistete empirische Forschungsbeitrag, die *Meta-Evaluation von IntuiBeat-F*, eine Reihe von Limitationen auf, denen sich zukünftige Forschung widmen sollte. Die wichtigsten der in Abschnitt 6.5 genannten und im Detail besprochenen Forschungsanregungen sind im Folgenden abschließend kurz zusammengefasst:

- Meta-Evaluation von IntuiBeat-F mit empirischem Nachweis des wissenschaftlichen Hauptgütekriteriums der Zuverlässigkeit durch Maße der Beurteilerübereinstimmung (siehe Hartson et al., 2001).
- Meta-Evaluation von IntuiBeat-F unter Berücksichtigung weiterer Nebengütekriterien (z.B. Downstream Utility, Praxistauglichkeit, Überzeugungskraft) zur Erfassung weiterer Aspekte praktischer Güte (siehe Blandford et al., 2008).
- Meta-Evaluation von IntuiBeat-F unter Berücksichtigung weiterer Untersuchungsgegenstände neben CUIs und Webseiten (z.B. Medizinprodukte), um eine Generalisierung der Ergebnisse auf weitere Untersuchungsgegenstände zu erlauben und der in der HCI vorherrschenden Übergeneralisierung von Methoden vorzubeugen (siehe Teilabschnitt 3.2.5).
- Meta-Evaluation von IntuiBeat-F mit Einsatz von Experten (d.h. Praktiker mit Berufserfahrung) anstelle von Novizen (d.h. Studierende mit Vorerfahrung bei der Nutzung von CUIs) als Anwender von IntuiBeat-F zur Generalisierung der Ergebnisse auf Evaluatoren mit höherer Vorerfahrung bei der Ausführung von Nutzertests und Domänenkenntnis (z.B. Vorerfahrung mit getesteten/ähnlichen CUIs).
- Meta-Evaluation von IntuiBeat-F mit Einsatz von Experten (z.B. Praktikern mit Vorerfahrung in bestimmten Domänen wie getesteten/ähnlichen CUIs oder getesteten/ähnlichen Medizinprodukten) anstelle von Novizen (d.h. Studierende mit Vorerfahrung in bestimmter Domäne) als mit der Methode getestete Nutzergruppe zur Generalisierung der Ergebnisse auf andere Nutzergruppen.

7.7 Schlussfolgerung und Ausblick

Kann die intuitive Benutzung von User Interfaces (speziell 3D-CUIs) sowohl formativ als auch summativ mit möglichst hoher zeitlicher Anwendungseffizienz mithilfe der im Rahmen dieser Dissertation entwickelten Methoden IntuiBeat-S und IntuiBeat-F evaluiert werden? Unter Berücksichtigung der in dieser Arbeit dargestellten Theorie und der Ergebnisse der durchgeführten Experimente, kann diese Frage mit „ja“ beantwortet werden. Die ersten drei Experimente konnten dabei die wissenschaftliche Güte und die zeitliche Anwendungseffizienz von IntuiBeat-S im Vergleich zu anderen, in der HCI verfügbaren, summativen Evaluationsmethoden bei der Evaluation von verschiedenen 3D-CUI-Interaktionslösungen demonstrieren (siehe Kapitel 5). Durch die folgenden vier Experimente konnten daraufhin die wissenschaftliche Güte und die zeitliche Anwendungseffizienz von IntuiBeat-F im Vergleich zum Nutzertest mit retrospektivem Think-Aloud-Protokoll bei der Evaluation von 3D-CUI-Interaktionslösungen gezeigt werden (siehe Kapitel 6). Wie bereits in den letzten Abschnitten zusammengefasst wurde, konnte dieser Arbeit darüber hinaus noch weitere Forschungsbeiträge für die HCI erbringen. Insbesondere wurde durch die Ableitung einer Messdefinition aus der metatheoretischen Betrachtung der Definitionen, der im Forschungsfeld tätigen drei Hauptforschergruppen, ein für die Evaluation intuitiver Benutzung geeigneter theoretischer Rahmen geschaffen, auf dessen Grundlage intuitive Benutzung mithilfe verschiedener Evaluationsmethoden „messbar“ gemacht werden kann.

Jedoch konnten die im Rahmen dieser Dissertation durchgeführten Experimente die wissenschaftliche Güte und mit der zeitlichen Anwendungseffizienz nur einen Teilaspekt praktischer Güte im Bereich von CUIs demonstrieren. Im Rahmen der in den letzten Abschnitten diskutierten Forschungsbeiträge, wurden eine Reihe von Ansatzpunkten für künftige Forschung gegeben. Diese Ansatzpunkte lassen sich in Form von vier Bereichen nach der Art des geleisteten Forschungsbeitrags zusammenfassen, wobei hier einfachheitshalber der methodische und empirische Forschungsbeitrag als ein Bereich betrachtet wird.

Übersichtsarbeiten Die im Rahmen dieser Dissertation erbrachten Übersichten über vorhandene Definitionen intuitiver Benutzung und bereits vorhandene Evaluationsmethoden für intuitive Benutzung erfolgten lediglich narrativ. Auch frühere Übersichtsarbeiten fokussierten sich ausschließlich auf diese Art von Übersichtsarbeit. Daher sollten sich zukünftige Forschungsarbeiten auch systematischen Übersichtsarbeiten widmen, um bereits bestehende Übersichtsarbeiten ergänzen zu können.

Theoretischer Forschungsbeitrag Des Weiteren konnte im Rahmen dieser Dissertation erstmalig eine Messdefinition für intuitive Benutzung vorgestellt werden, die es erlaubt, das Ausmaß intuitiver Benutzung anhand der mentalen Beanspruchung als objektives Merkmal, des Gefühls von Flüssigkeit als subjektives Merkmal und der Effektivität als pragmatisches Merkmal zu beurteilen. Zukünftige Forschung in der HCI sollte bei der Beurteilung intuitiver Benutzung die Phänomenologie, der mit dem Gefühl von Flüssigkeit assoziierten subjektiven Gefühle weiter untersuchen. Zudem sollte die mentale mentale Beanspruchung anhand von mathematischen Modellen genauer spezifiziert und diese Modelle zur Vorhersage mentaler Beanspruchung in verschiedenen Nutzungskontexten genutzt werden. Schließlich sollten sowohl das Gefühl von Flüssigkeit als auch die mentale Beanspruchung in Form von Traits bei der Evaluation intuitiver Benutzung berücksichtigt werden.

Forschungsbeitrag in Form eines Werkzeugs Im Gegensatz zu bestehenden, objektiven Evaluationsmethoden wurde zur Standardisierung der Aufzeichnung und Auswertung von Rhythmusabweichungen im Zuge dieser Dissertation ein Setup, bestehend aus Hard- und Software, entwickelt. Um die Aufzeichnung und formative oder summative Auswertung von Rhythmusabweichungen für den Evaluator noch weiter zu erleichtern, sollte sich zukünftige Forschung damit beschäftigen, die Aufzeichnung mithilfe eines direkt am Fuß angebrachten Bewegungssensors zu ermöglichen, um so eine zuverlässige Aufzeichnung unabhängig von der Schuhform zuzulassen. Ferner kann unter anderem der Zusammenschluss der IntuiBeat-Software und der IntuiBeat-Analyse-Software die Arbeit des Evaluators erleichtern und flexibler machen.

Methodische und empirische Forschungsbeiträge Obwohl anhand der im Rahmen dieser Dissertation durchgeführten Experimente sowohl die wissenschaftliche Güte als auch die zeitliche Anwendungseffizienz von IntuiBeat-F und IntuiBeat-S demonstriert werden konnten, sollten sich zukünftige Arbeiten auch auf andere Aspekte praktischer Güte konzentrieren und versuchen, die im CUI-Bereich gesammelten Ergebnisse auch unter Berücksichtigung anderer Nutzergruppen (z.B. Praktikern mit Berufserfahrung), anderer Domänen (z.B. Medizinprodukte), anderer Aufgaben und anderer Anwender der Methoden mit höherer Vorerfahrung bei der Durchführung von Nutzertests zu replizieren.

Die formative und summative Evaluation intuitiver Benutzung ist noch ein junges Forschungsgebiet. Um sicherzustellen, dass trotz der fortschreitenden Digitalisierung innerhalb der Gesellschaft auch Menschen mit unterschiedlicher Vorerfahrung verschiedene technische Systeme intuitiv benutzen können, wird sie jedoch einen immer größeren Stellenwert in der Mensch-Maschine-Interaktion erhalten. Diese Arbeit hat zu diesem Zweck sowohl eine formative (d.h. IntuiBeat-F) als auch eine summative (d.h. IntuiBeat-S) Evaluationsmethode für intuitive Benutzung vorgeschlagen und deren Güte (d.h. wissenschaftliche Güte und zeitliche Anwendungseffizienz) im Bereich von 3D-Anwendungen, die im Zuge der Digitalisierung für alle Nutzergruppen bedeutsamer werden (siehe Kapitel 1), nachgewiesen. Die Ergebnisse dieser Arbeit zeigen, dass beide Methoden für die Evaluation intuitiver Benutzung vielversprechend sind und breit eingesetzt werden können. Mit der Hilfe dieser neuen Methoden kann die intuitive Benutzung einer Vielzahl von User Interfaces (nicht nur CUIs oder Webseiten) zeitlich effizient optimiert (d.h. durch IntuiBeat-F) und bewertet (d.h. durch InuiBeat-S) werden. Auf diese Weise können Wissenschaftler und Praktiker gleichermaßen bei der Entwicklung intuitiver, technischer Systeme, die von Menschen mit unterschiedlicher Vorerfahrung genutzt werden können, unterstützt werden und damit zu einer inklusiven Gesellschaft in Zeiten der Digitalisierung beitragen.

Literatur

Aasman, J., Mulder, G., & Mulder, L. J. (1987). Operator effort and the measurement of heart-rate variability. *Human Factors*, *29*(2), 161–170.

Ackerman, R., & Thompson, V. (2015). Meta-reasoning. In A. Feeney & V. Thompson (Hrsg.), *Reasoning as Memory* (S. 164–182). Hove, Vereinigtes Königreich: Psychology Press.

Ackerman, R., & Thompson, V. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*(8), 607–617.

Adams, C. E., & Leverland, M. B. (1985). Environmental and behavioral factors that can affect blood pressure. *The Nurse Practitioner*, *10*(11), 39–40.

Adler, R. F., & Benbunan-Fich, R. (2015). The effects of task difficulty and multitasking on performance. *Interacting with Computers*, *27*(4), 430–439.

Agarwal, R., & Venkatesh, V. (2002). Assessing a firm's web presence: A heuristic evaluation procedure for the measurement of usability. *Information Systems Research*, *13*(2), 168–186.

Agor, W. H. (1986). *The logic of intuitive decision making: A research-based approach for top management*. New York, NY, Vereinigte Staaten: Quorum Books.

Akers, D. L. (2010). *Backtracking events as indicators of software usability problems* (Unveröffentlichte Dissertation), Stanford University, Standford, CA, Vereinigte Staaten.

Akers, D. L., Jeffries, R., Simpson, M., & Winograd, T. (2012). Backtracking events as indicators of usability problems in creation-oriented applications. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *19*(2), 16.

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187–195.

Albers, M. J. (2011). Tapping as a measure of cognitive load and website usability. In *Proceedings of the 29th ACM International Conference on Design of Communication* (S. 25–32). ACM. New York, NY, Vereinigte Staaten.

Alexander, C. (1977). *A pattern language: Towns, buildings, construction*. Oxford, Vereinigtes Königreich: Oxford University Press.

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, *28*(3), 1–39.

Allport, D. A. (1980). Attention and performance. *Cognitive Psychology: New Directions*, *1*, 12–153.

Alshamari, M., & Mayhew, P. (2009). Technical review: Current issues of usability testing. *IETE Technical Review*, *26*(6), 402–406.

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology*, *136*(4), 569.

Anderson, J. R. (2013). *The architecture of cognition*. London, Vereinigtes Königreich: Psychology Press.

Anderson, J. R., & Funke, J. (2001). *Kognitive Psychologie*. Heidelberg, Deutschland: Spektrum Akademischer Verlag.

Antle, A. N., Corness, G., & Droumeva, M. (2009). Human-computer-intuition? Exploring the cognitive basis for intuition in embodied interaction. *International Journal of Arts and Technology*, *2*(3), 235–254.

Apparies, R. J., Riniolo, T. C., & Porges, S. W. (1998). A psychophysiological investigation of the effects of driving longer-combination vehicles. *Ergonomics*, *41*(5), 581–592.

Asikhia, O. K. (2015). *Evaluating intuitive interactions using image schemas* (Unveröffentlichte Dissertation), Cardiff University, Cardiff, Vereinigtes Königreich.

Astor, M., Jarowinsky, M., & von Lukas, U. (2013). *Marktperspektiven von 3D in industriellen Anwendungen*. Berlin, Deutschland: Bundesministerium für Wirtschaft und Technologie.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In *Psychology of Learning and Motivation* (Bd. 2, S. 89–195). Amsterdam, Niederlande: Elsevier.

Auf der Heide, B. H. (1993). Welche software-ergonomischen Evaluationsverfahren können was leisten? In K. H. Rödiger (Hrsg.), *Software-Ergonomie'93* (S. 157–171). Heidelberg, Deutschland: Springer.

Autodesk. (2017a). Fusion 360 (Version 2.9.3706) [Software]. Abgerufen von https://www.autodesk.de/products/fusion-360/.

Autodesk. (2017b). Tinkercad (Version 3.6) [Software]. Abgerufen von https://www.tinkercad.com/.

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction*, *16*(5), 389–400.

Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators & Virtual Environments*, *6*(4), 355–385.

Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge, Vereinigtes Königreich: Cambridge University Press.

Backs, R. W. (1995). Going beyond heart rate: Autonomic space and cardiovascular assessment of mental workload. *The International Journal of Aviation Psychology*, *5*(1), 25–48.

Backs, R. W., Navidzadeh, H. T., & Xu, X. (2000). Cardiorespiratory indices of mental workload during simulated air traffic control. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Bd. 44, *13*, S. 89–92). Sage Publications. Los Angeles, CA, Vereinigte Staaten.

Backs, R. W., & Seljos, K. A. (1994). Metabolic and cardiorespiratory measures of mental effort: The effects of level of difficulty in a working memory task. *International Journal of Psychophysiology*, *16*(1), 57–68.

Baddeley, A. (1966). The capacity for generating information by randomization. *The Quarterly Journal of Experimental Psychology*, *18*(2), 119–129.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556.

Baddeley, A. (1999). Working memory: The multiple component model. In *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (S. 28–61). Cambridge, Vereinigtes Königreich: Cambridge University Press.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Baddeley, A. (2001). The magic number and the episodic buffer. *Behavioral and Brain Sciences*, *24*(1), 117–118.

Baddeley, A. (2007). *Working memory, thought, and action.* Oxford, Vereinigtes Königreich: Oxford University Press.

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29.

Baddeley, A., & Hitch, G. (1974). Working memory. In *Psychology of Learning and Motivation* (Bd. 8, S. 47–89). Amsterdam, Niederlande: Elsevier.

Bailey, B. P., & Iqbal, S. T. (2008). Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *14*(4), 21.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human–Computer Interaction*, *24*(6), 574–594.

Barendregt, W., Bekker, M. M., Bouwhuis, D. G., & Baauw, E. (2006). Identifying usability and fun problems in a computer game during first use and after some practice. *International Journal of Human-Computer Studies*, *64*(9), 830–846.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, *54*(7), 462.

Bargh, J. A., & Gollwitzer, P. M. (1994). Environmental control of goal-directed action: Automatic and strategic contingencies between situations and behavior. In *Nebraska Symposium on Motivation. Nebraska Symposium on Motivation* (Bd. 41, S. 71–124). University of Nebraska Press. Lincoln, NE, Vereinigte Staaten.

Barnum, C. M. (2010). *Usability testing essentials: Ready, set... test!* Amsterdam, Niederlande: Elsevier.

Barrouillet, P., Portrat, S., & Camos, V. (2011). On the law relating processing to storage in working memory. *Psychological Review*, *118*(2), 175.

Basso, D., & Belardinelli, M. O. (2006). The role of the feedforward paradigm in cognitive psychology. *Cognitive Processing*, *7*(2), 73–88.

Bastick, T. (1982). *Intuition: How we think and act.* Hoboken, NJ, Vereinigte Staaten: John Wiley & Sons.

Bastick, T. (2003). *Intuition: Evaluating the construct and its impact on creative thinking.* New York, NY, Vereinigte Staaten: Stoneman & Lang.

Baumeister, R. F., & Leary, M. R. (1997). Writing narrative literature reviews. *Review of General Psychology*, *1*(3), 311–320.

Baumgartner, P. (1999). Evaluation mediengestützten Lernens. Theorie-Logik-Modelle. In M. Kindt (Hrsg.), *Projektevaluation in der Lehre - Multimedia an Hochschulen zeigt Profile* (S. 61–97). Münster, Deutschland: Waxmann Verlag.

Bazerman, M. H., Tenbrunsel, A. E., & Wade-Benzoni, K. (1998). Negotiating with yourself and losing: Making decisions with competing internal preferences. *Academy of Management Review*, *23*(2), 225–241.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276.

Beauducel, A., & Leue, A. (2014). *Psychologische Diagnostik*. Göttingen, Deutschland: Hogrefe Verlag.

Bergman, O., Tene-Rubinstein, M., & Shalom, J. (2013). The use of attention resources in navigation versus search. *Personal and Ubiquitous Computing*, *17*(3), 583–590.

Berkovits, I., Hancock, G. R., & Nevitt, J. (2000). Bootstrap resampling approaches for repeated measure designs: Relative robustness to sphericity and normality violations. *Educational and Psychological Measurement*, *60*(6), 877–892.

Bernardi, L., Wdowczyk-Szulc, J., Valenti, C., Castoldi, S., Passino, C., Spadacini, G., & Sleight, P. (2000). Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability. *Journal of the American College of Cardiology*, *35*(6), 1462–1469.

Berry, W. D., Feldman, S., & Stanley Feldman, D. (1985). *Multiple regression in practice*. Thousand Oaks, CA, Vereinigte Staaten: Sage Publications.

Berti, S. (2008). Cognitive control after distraction: Event-related brain potentials (ERPs) dissociate between different processes of attentional allocation. *Psychophysiology*, *45*(4), 608–620.

Betsch, T. (2008). The nature of intuition and its neglect in research on judgment and decision making. *Intuition in Judgment and Decision Making*, 3–22.

Betsch, T., Funke, J., & Plessner, H. (2011). *Allgemeine Psychologie für Bachelor: Denken-Urteilen, Entscheiden, Problemlösen.* Heidelberg, Deutschland: Springer.

Betsch, T., & Glöckner, A. (2010). Intuition in judgment and decision making: Extensive thinking without effort. *Psychological Inquiry*, *21*(4), 279–294.

Beyer, H., & Holtzblatt, K. (1997). *Contextual design: Defining customer-centered systems*. Amsterdam, Niederlande: Elsevier.

Bias, R. G., Kortum, P., Sauro, J., & Gillan, D. (2013). Clothing the naked emperor: The unfulfilled promise of the science of usability. *Interactions*, *20*(6), 72–77.

Bickerton, D. (1995). Language and human behavior. In *Based on three public lectures presented at University Washington, Seattle, Oct 1992.* University of Washington Press, Seattle, WA, Vereinigte Staaten.

Blackler, A. (2006). *Intuitive interaction with complex artefacts* (Unveröffentlichte Dissertation), Queensland University of Technology, Brisbane, Australien.

Blackler, A. (2008). *Intuitive interaction with complex artefacts: Empirically-based research*. Saarbrücken, Deutschland: VDM Verlag Dr. Müller.

Blackler, A. (2018). *Intuitive Interaction: Research and Application*. Boca Raton, FL, Vereinigte Staaten: CRC Press.

Blackler, A., & Hurtienne, J. (2007). Towards a unified view of intuitive interaction: Definitions, models and tools across the world. *MMI-Interaktiv*, *13*(2007), 36–54.

Blackler, A., & Popovic, V. (2015). Towards intuitive interaction theory. *Interacting with Computers*, *27*(3), 203–209.

Blackler, A., Popovic, V., & Desai, S. (2018). Research Methods for Intuitive Interaction. In *Intuitive Interaction* (S. 65–88). Boca Raton, FL, Vereinigte Staaten: CRC Press.

Blackler, A., Popovic, V., Lawry, S., Reddy, R., Mahar, D. P., Kraal, B., & Chamorro-Koc, M. (2011). Researching intuitive interaction. In *Proceedings of the 4th IASDR (The International Association of Societies of Design Research Congress), IASDR 2011*. IASDR, Delft, Niederlande.

Blackler, A., Popovic, V., & Mahar, D. (2010). Investigating users' intuitive interaction with complex artefacts. *Applied Ergonomics*, *41*(1), 72–92.

Blackler, A., Popovic, V., & Mahar, D. P. (2002). Intuitive use of products. In *Design Research Society (DSR) International Conference: Common Ground* (S. 1–15). Staffordshire University Press. Staffordshire, Vereinigtes Königreich.

Blackler, A., Popovic, V., & Mahar, D. P. (2003). Designing for Intuitive Use of Products: An Investigation. In T. Yamanaka, M. Kubo, & K. Sato (Hrsg.), *Asian Design International Conference* (S. 1–16). Tsukuba, Japan.

Blackler, A., Popovic, V., & Mahar, D. P. (2005). Intuitive Interaction Applied to Interface Design. In *Proceedings of the 1st IASDR (The International Association of Societies of Design Research Congress), IASDR 2005*. IASDR, Douliou, Taiwan.

Blanca, M., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, *29*(4), 552–557.

Blandford, A., Green, T. R., Furniss, D., & Makri, S. (2008). Evaluating system utility and conceptual fit using CASSM. *International Journal of Human-Computer Studies*, *66*(6), 393–409.

Blettner, M., Sauerbrei, W., Schlehofer, B., Scheuchenpflug, T., & Friedenreich, C. (1997). Vergleich von traditionellen Reviews, Metaanalysen und gepoolten Analysen zur Bewertung von Risikofaktoren. *Informatik Biometrie und Epidemiologie in Medizin und Biologie*, *28*(3), 148–166.

Bodala, I. P., Ke, Y., Mir, H., Thakor, N. V., & Al-Nashash, H. (2014). Cognitive workload estimation due to vague visual stimuli using saccadic eye movements. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (S. 2993–2996). IEEE. Piscataway, NJ, Vereinigte Staaten.

Boles, D. B., Bursk, J. H., Phillips, J. B., & Perdelwitz, J. R. (2007). Predicting dual-task performance with the Multiple Resources Questionnaire (MRQ). *Human Factors*, *49*(1), 32–45.

Bongers, K. C., Dijksterhuis, A., & Spears, R. (2010). On the role of consciousness in goal pursuit. *Social Cognition*, *28*(2), 262–272.

Borchers, J. O., & Thomas, J. C. (2001). Patterns: What's in it for HCI? In *CHI'01 Extended Abstracts on Human Factors in Computing Systems* (S. 225–226). New York, NY, Vereinigte Staaten: ACM.

Borsci, S., Macredie, R. D., Barnett, J., Martin, J., Kuljis, J., & Young, T. (2013). Reviewing and extending the five-user assumption: A grounded procedure for interaction evaluation. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *20*(5), 29.

Bortz, J., & Schuster, C. (2011). *Statistik für Human-und Sozialwissenschaftler: Limitierte Sonderausgabe*. Heidelberg, Deutschland: Springer.

Bouma, H. (1962). Size of the static pupil as a function of wave-length and luminosity of the light incident on the human eye. *Nature*, *193*(4816), 690–691.

Bousefsaf, F., Maaoui, C., & Pruski, A. (2014). Remote detection of mental workload changes using cardiac parameters assessed with a low-cost webcam. *Computers in Biology and Medicine*, *53*, 154–163.

Boutcher, Y. N., & Boutcher, S. H. (2006). Cardiovascular response to Stroop: Effect of verbal response and task difficulty. *Biological Psychology*, *73*(3), 235–241.

Bowers, K. S., Farvolden, P., & Mermigis, L. (1995). Intuitive antecedents of insight. In *The Creative Cognition Approach* (S. 27–51). Cambridge, MA, Vereinigte Staaten: Bradford Books.

Bowers, V. A., & Snyder, H. L. (1990). Concurrent versus retrospective verbal protocol for comparing window usability. In *Proceedings of the Human Factors Society Annual Meeting* (Bd. 34, S. 1270–1274). Los Angeles, CA, Vereinigte Staaten: Sage Publications.

Braby, C., Harris, D., & Muir, H. (1993). A psychophysiological approach to the assessment of work underload. *Ergonomics*, *36*(9), 1035–1042.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal*, *5*(4), 437–474.

Brainerd, C. J., & Reyna, V. F. (2001). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. *Advances in Child Development and Behavior*, *28*(28), 41–100.

Brandenburg, S., & Sachse, K. (2012). Intuition comes with experience. In *Proceedings of the Human Factors and Ergonomics Society of Europe Conference* (S. 213–223). Groningen, Niederlande: HFES Europe.

Brewer, W. F., & Nakamura, G. V. (1984). *The nature and functions of schemas* (Techn. Ber. Nr. 325). University of Illinois. Chicago, IL, Vereinigte Staaten.

Brooke, J. (1996). SUS - A quick and dirty usability scale. *Usability Evaluation in Industry*, *189*(194), 4–7.

Brookhuis, K. A., & De Waard, D. (1993). The use of psychophysiology to assess driver status. *Ergonomics*, *36*(9), 1099–1110.

Brookhuis, K. A., & De Waard, D. (2010). Monitoring drivers' mental workload in driving simulators using physiological measures. *Accident Analysis & Prevention*, *42*(3), 898–903.

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology*, *42*(3), 361–377.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin*, *109*(2), 204.

Brown, R. G., & Marsden, C. D. (1991). Dual task performance and processing resources in normal subjects and patients with Parkinson's disease. *Brain*, *114*(1), 215–231.

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 53–61.

Brünken, R., Plass, J. L., & Leutner, D. (2004). Assessment of cognitive load in multimedia learning with dual-task methodology: Auditory load and modality effects. *Instructional Science*, *32*(1-2), 115–132.

Bühner, M. (2011). *Einführung in die Test-und Fragebogenkonstruktion.* Hallbergmoos, Deutschland: Pearson Studium.

Burmester, M. (2016). *Auswertung des Usability Tests.* Hochschule der Medien Stuttgart. Stuttgart, Deutschland.

Burmester, M., Haasler, K., Schippert, K., Engel, V., Tille, R., Reinhardt, D., & Hurtienne, J. (2018). Lost in Space? 3D-Interaction-Patterns für einfache und positive Nutzung von 3D Interfaces. In S. Hess & H. Fischer (Hrsg.), *Mensch und Computer 2018 - Usability Professionals* (S. 53–66). Bonn, Deutschland: Gesellschaft für Informatik e. V. und German UPA e. V.

Byers, J. C. (1989). Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary? *Advances in Industrial Ergonomics and Safety*, 481–485.

Cain, B. (2007). *A review of the mental workload literature* (Techn. Ber. Nr. RTO-TRHFM-121-Part-II). Defence Research and Development Canada Toronto. Human System Integration Section.

Caine, K. (2016). Local standards for sample size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (S. 981–992). New York, NY, Vereinigte Staaten: ACM.

Caldwell, J. A., Wilson, G. F., & Cetinguc, M. (1994). *Psychophysiological assessment methods.* Paris, Frankreich: AGARD.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81.

Capra, M. G. (2002). Contemporaneous versus retrospective user-reported critical incidents in usability evaluation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Bd. 46, *24*, S. 1973–1977). Sage Publications. Los Angeles, CA, Vereinigte Staaten.

Card, S. K., Moran, T. P., & Newell, A. (1980). The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, *23*(7), 396–410.

Casasanto, D., & Chrysikou, E. G. (2011). When left is "right" motor fluency shapes abstract concepts. *Psychological Science*, *22*(4), 419–422.

Causse, M., Sénard, J.-M., Démonet, J. F., & Pastor, J. (2010). Monitoring cognitive and emotional processes through pupil and cardiac response during dynamic versus logical task. *Applied Psychophysiology and Biofeedback*, *35*(2), 115–123.

Chaiken, S. (1987). The heuristic model of persuasion. In *Social Influence: The Ontario Symposium* (Bd. 5, S. 3–39). Hillsdale, NJ, Vereinigte Staaten: Lawrence Erlbaum.

Chaiken, S., & Trope, Y. (1999). *Dual-process theories in social psychology.* New York, NY, Vereinigte Staaten: Guilford Press.

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: a systematic review. *Applied Ergonomics*, *74*, 221–232.

Chartrand, T. L., & Bargh, J. A. (1996). Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of Personality and Social Psychology*, *71*(3), 464.

Chattopadhyay, D., & Bolchini, D. (2014). Touchless circular menus: toward an intuitive UI for touchless interactions with large displays. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces* (S. 33–40). New York, NY, Vereinigte Staaten: ACM.

Chattratichart, J., & Brodie, J. (2004). Applying user testing data to UEM performance metrics. In *CHI'04 Extended Abstracts on Human Factors in Computing Systems* (S. 1119–1122). New York, NY, Vereinigte Staaten: ACM.

Chefkoch GmbH. (2017). Chefkoch.de. Abgerufen von https://www.chefkoch.de/.

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., & Conway, D. (2016). *Robust multimodal cognitive load measurement.* Heidelberg, Deutschland: Springer.

Chen, S. [Serena], & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In *Dual-process Theories in Social Psychology.* (S. 73–96). New York, NY, Vereinigte Staaten: Guilford Press.

Chen, S. [Siyuan], & Epps, J. (2013). Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer Methods and Programs in Biomedicine*, *110*(2), 111–124.

Chen, S. [Siyuan], & Epps, J. (2014). Using task-induced pupil diameter and blink rate to infer cognitive load. *Human–Computer Interaction*, *29*(4), 390–413.

Cheng, J., Okoso, A., Kunze, K., Henze, N., Schmidt, A., Lukowicz, P., & Kise, K. (2014). On the tip of my tongue: A non-invasive pressure-based tongue interface. In *Proceedings of the 5th Augmented Human International Conference* (S. 12). New York, NY, Vereinigte Staaten: ACM.

Chi, E. H., Rosien, A., Supattanasiri, G., Williams, A., Royer, C., Chow, C., ... Cousins, S. (2003). The bloodhound project: Automating discovery of web usability issues using the InfoScent simulator. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (S. 505–512). New York, NY, Vereinigte Staaten: ACM.

Chopine, A. (2012). *3D art essentials*. Abingdon, Vereinigtes Königreich: Routledge.

Cienki, A. (2015). Image Schemas and Mimetic Schemas in Cognitive Linguistics and Gesture Studies. *Multimodality and Cognitive Linguistics*, *78*, 195.

Cienki, A., & Müller, C. (2008). *Metaphor and gesture*. Amsterdam, Niederlande: John Benjamins Publishing.

Clark, R. C., & Mayer, R. E. (2016). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. Hoboken, NJ, Vereinigte Staaten: John Wiley & Sons.

Clearleft Ltd. (2016). Silverback (Version 3.1.4) [Software]. Abgerufen von https://silverbackapp.com/.

Clore, G. L., & Huntsinger, J. R. (2007). How emotions inform judgment and regulate thought. *Trends in Cognitive Sciences*, *11*(9), 393–399.

Cockburn, A., Quinn, P., & Gutwin, C. (2015). Examining the peak-end effects of subjective experience. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (S. 357–366). New York, NY, Vereinigte Staaten: ACM.

Cohen, G., & Martin, M. (1975). Hemisphere differences in an auditory Stroop test. *Perception & Psychophysics*, *17*(1), 79–83.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Abingdon, Vereinigtes Königreich: Routledge.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155.

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature*, *386*(6625), 604.

Cohen, M. A., Cavanagh, P., Chun, M. M., & Nakayama, K. (2012). The attentional requirements of consciousness. *Trends in Cognitive Sciences*, *16*(8), 411–417.

Collet, C., Clarion, A., Morel, M., Chapon, A., & Petit, C. (2009). Physiological and behavioural changes associated to the management of secondary tasks while driving. *Applied Ergonomics*, *40*(6), 1041–1046.

Cooper, G. E., & Harper, R. P. J. (1969). *The use of pilot rating in the evaluation of aircraft handling qualities* (Techn. Ber. Nr. NASA TN-D-5153). National Aeronautics and Space Administration. Washington DC, Vereinigte Staaten.

Cooper, R., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*(4), 297–338.

Cooper, R., & Shallice, T. (2006). Hierarchical schemas and goals in the control of sequential behavior. *Psychological Review*, *113*(4), 887–916.

Courtney, A. J., & Shou, C. H. (1985). Simple measures of visual-lobe size and search performance. *Ergonomics*, *28*(9), 1319–1331.

Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: A review. *International Journal of Psychological Research*, *3*(1), 58–67.

Cowley, B., Filetti, M., Lukander, K., Torniainen, J., Henelius, A., Ahonen, L., ... Huotilainen, M. (2016). The psychophysiology primer: A guide to methods and a broad review with a focus on human–computer interaction. *Foundations and Trends in Human–Computer Interaction*, *9*(3), 151–308.

Cramer, E. M., & Bock, R. D. (1966). Chapter VIII: Multivariate analysis. *Review of Educational Research*, *36*(5), 604–617.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Crumlish, C., & Malone, E. (2009). *Designing social interfaces: Principles, patterns, and practices for improving the user experience.* Newton, MA, Vereinigte Staaten: O'Reilly.

Cruz, A. L. F., Arango-Muñoz, S., & Volz, K. G. (2016). Oops, scratch that! Monitoring one's own errors during mental calculation. *Cognition*, *146*, 110–120.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain.* New York, NY: HarperCollins.

Damos, D. (1985). The effect of asymmetric transfer and speech technology on dual-task performance. *Human Factors*, *27*(4), 409–421.

Damos, D. (1991). *Multiple task performance.* Boca Raton, FL, Vereinigte Staaten: CRC Press.

Dawson, M. E., Schell, A. M., & Filion, D. L. (2017). *The electrodermal system.* Cambridge, Vereinigtes Königreich: Cambridge University Press.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*(1), 1–18.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PloS One*, *6*(1), e15954.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299.

De Rivecourt, M., Kuperus, M., Post, W., & Mulder, L. J. (2008). Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics*, *51*(9), 1295–1319.

De Waard, D. (1996). *The measurement of drivers' mental workload.* Groningen, Niederlande: Groningen University, Traffic Research Center Netherlands.

De Winter, J. C. (2014). Controversy in human factors constructs and the explosive use of the NASA-TLX: A measurement perspective. *Cognition, Technology & Work*, *16*(3), 289–297.

Dearden, A., & Finlay, J. (2006). Pattern languages in HCI: A critical review. *Human–Computer Interaction*, *21*(1), 49–102.

DeGEval. (2016). *Standards für Evaluation.* Mainz, Deutschland: DeGEval – Gesellschaft für Evaluation.

Dehais, F., Causse, M., Vachon, F., & Tremblay, S. (2012). Cognitive conflict in human–automation interactions: a psychophysiological study. *Applied Ergonomics*, *43*(3), 588–595.

Del Galdo, E. M., Williges, R. C., Williges, B. H., & Wixon, D. R. (1986). An evaluation of critical incidents for software documentation design. In *Proceedings of the Human Factors Society Annual Meeting* (Bd. 30, *1*, S. 19–23). Sage Publications. Los Angeles, CA, Vereinigte Staaten.

Delaney, J., & Brodie, D. (2000). Effects of short-term psychological stress on the time and frequency domains of heart-rate variability. *Perceptual and Motor Skills*, *91*(2), 515–524.

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology*, *100*(1), 223.

Dell, N., Vaidyanathan, V., Medhi, I., Cutrell, E., & Thies, W. (2012). Yours is better!: Participant response bias in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (S. 1321–1330). ACM. New York, NY, Vereinigte Staaten.

Demberg, V. (2013). Pupillometry: The index of cognitive activity in a dual-task study. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Bd. 35, *35*), Austin, TX, Vereinigte Staaten.

Derish, P. A., & Annesley, T. M. (2011). How to write a rave review. *Clinical Chemistry*, *57*(3), 388–391.

Desai, S., Blackler, A., & Popovic, V. (2015). Intuitive use of tangible toys. In *Proceedings of the 6th IASDR (The International Association of Societies of Design Research Congress), IASDR 2015* (S. 522–540). IASDR. Brisbane, Australien.

Dey, A., & Mann, D. D. (2010). Sensitivity and diagnosticity of NASA-TLX and simplified SWAT to assess the mental workload associated with operating an agricultural sprayer. *Ergonomics*, *53*(7), 848–857.

Di Nocera, F., Camilli, M., & Terenzi, M. (2007). A random glance at the flight deck: Pilots' scanning strategies and the real-time assessment of mental workload. *Journal of Cognitive Engineering and Decision Making*, *1*(3), 271–285.

Di Nocera, F., Terenzi, M., & Camilli, M. (2006). Another look at scanpath: Distance to nearest neighbour as a measure of mental workload. *Developments in Human Factors in Transportation, Design, and Evaluation*, 295–303.

Di Stasi, L. L., Renner, R., Staehr, P., Helmert, J. R., Velichkovsky, B. M., Cañas, J. J., ... Pannasch, S. (2010). Saccadic peak velocity sensitivity to variations in mental workload. *Aviation, Space, and Environmental Medicine*, *81*(4), 413–417.

Diefenbach, S., & Hassenzahl, M. (2017). *Psychologie in der nutzerzentrierten Produktgestaltung.* Heidelberg, Deutschland: Springer.

Diefenbach, S., & Ullrich, D. (2015). An experience perspective on intuitive interaction: Central components and the special effect of domain transfer distance. *Interacting with Computers*, *27*(3), 210–234.

Dijksterhuis, A., & Aarts, H. (2010). Goals, attention, and (un) consciousness. *Annual Review of Psychology*, *61*, 467–490.

Dijksterhuis, A., & Van Olden, Z. (2006). On the benefits of thinking unconsciously: Unconscious thought can increase post-choice satisfaction. *Journal of Experimental Social Psychology*, *42*(5), 627–631.

Dochy, F., & Alexander, P. A. (1995). Mapping prior knowledge: A framework for discussion among researchers. *European Journal of Psychology of Education*, *10*(3), 225–242.

Dochy, F., Segers, M., & Buehl, M. M. (1999). The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of Educational Research*, *69*(2), 145–186.

Dorfman, J., Shames, V. A., & Kihlstrom, J. F. (1996). Intuition, incubation, and insight: Implicit cognition in problem solving. In *Implicit Cognition* (S. 257–296). Oxford, Vereinigtes Königreich: Oxford University Press.

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation*. Heidelberg, Deutschland: Springer.

Drory, A. (1985). Effects of rest and secondary task on simulated truck-driving task performance. *Human Factors*, *27*(2), 201–207.

Drugge, M., & Witt, H. (2006). Hotwire: An apparatus for simulating primary tasks in wearable computing. In *ACM International Conference on Human Factors in Computing Systems* (S. 1535–1540). ACM. New York, NY, Vereinigte Staaten.

Dumas, J. S., Dumas, J. S., & Redish, J. (1999). *A practical guide to usability testing*. Bristol, Vereinigtes Königreich: Intellect books.

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Thousand Oaks, CA, Vereinigte Staaten: Sage Publications.

Đurić-Jovičić, M., Jovičić, N., Radovanović, S., Ječmenica-Lukić, M., Belić, M., Popović, M., & Kostić, V. (2018). Finger and foot tapping sensor system for objective motor assessment. *Vojnosanitetski pregled*, *75*(1), 68–77.

Dussault, C., Jouanin, J.-C., Philippe, M., & Guezennec, C.-Y. (2005). EEG and ECG changes during simulator operation reflect mental workload and vigilance. *Aviation, Space, and Environmental Medicine*, *76*(4), 344–351.

Dutt, A., Johnson, H., & Johnson, P. (1994). Evaluating evaluation methods. In *Proceedings of the Conference on People and Computers IX* (S. 109–121). Cambridge University Press. Cambridge, Vereinigtes Königreich.

Edgell, S. E., & Noon, S. M. (1984). Effect of violation of normality on the t test of the correlation coefficient. *Psychological Bulletin*, *95*(3), 576.

Eggemeier, F. T. (1988). Properties of workload assessment techniques. In *Advances in Psychology* (Bd. 52, S. 41–62). Amsterdam, Niederland: Elsevier.

Eggemeier, F. T., Wilson, G., Kramer, A., & Damos, D. (1991). Workload assessment in multi-task environments. London, Vereinigtes Königreich: Taylor & Francis.

Eid, M., & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Göttingen, Deutschland: Hogrefe Verlag.

Eilers, K., Nachreiner, F., & Hänecke, K. (1986). Entwicklung und Überprüfung einer Skala zur Erfassung subjektiv erlebter Anstrengung. *Zeitschrift für Arbeitswissenschaft*, (4), 214–224.

Ekstrom, R. B., Dermen, D., & Harman, H. H. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ, Vereinigte Staaten: Educational Testing Service.

Emerson, M. J., & Miyake, A. (2003). The role of inner speech in task switching: A dual-task investigation. *Journal of Memory and Language*, *48*(1), 148–168.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American psychologist*, *49*(8), 709.

Epstein, S. (2010). Demystifying intuition: What it is, what it does, and how it does it. *Psychological Inquiry*, *21*(4), 295–312.

Evans, J. S. B. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology*, *75*(4), 451–468.

Evans, J. S. B. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking & Reasoning*, *4*(1), 45–110.

Evans, J. S. B. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, *13*(3), 378–395.

Evans, J. S. B. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement.* London, Vereinigtes Königreich: Psychology Press.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278.

Evans, J. S. B. (2009). How many dual-process theories do we need? One, two, or many? In *In two minds: Dual processes and Beyond.* (S. 33–54).

Evans, J. S. B. (2010). *Thinking twice: Two minds in one brain.* Oxford, Vereinigtes Königreich: Oxford University Press.

Evans, J. S. B. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review*, *31*(2-3), 86–102.

Evans, J. S. B., & Lynch, J. (1973). Matching bias in the selection task. *British Journal of Psychology*, *64*(3), 391–397.

Evans, J. S. B., & Over, D. E. (1996). *Rationality and reasoning.* London, Vereinigtes Königreich: Psychology Press.

Evans, J. S. B., & Stanovich, K. E. (2013). Theory and metatheory in the study of dual processing: Reply to comments. *Perspectives on Psychological Science*, *8*(3), 263–271.

Evans, J. S. B., & Wason, P. C. (1976). Rationalization in a reasoning task. *British Journal of Psychology*, *67*(4), 479–486.

Fairclough, S. H., & Venables, L. (2006). Prediction of subjective states from psychophysiology: A multivariate approach. *Biological Psychology*, *71*(1), 100–110.

Fairclough, S. H., Venables, L., & Tattersall, A. (2005). The influence of task demand and learning on the psychophysiological response. *International Journal of Psychophysiology*, *56*(2), 171–184.

Farmer, E., Belyavin, A., Jordan, C., Bunting, A., Tattersall, A., & Jones, D. (1995). *Predictive workload assessment: Final report.* Defence Research Agency. Farnborough, Vereinigtes Königreich.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160.

Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, *35*(3), 379–383.

Faure, V., Lobjois, R., & Benguigui, N. (2016). The effects of driving environment complexity and dual tasking on drivers' mental workload and eye blink behavior. *Transportation Research Part F: Traffic, Psychology and Behaviour*, *40*, 78–90.

Fazendeiro, T., Winkielman, P., Luo, C., & Lorah, C. (2005). False recognition across meaning, language, and stimulus format: Conceptual relatedness and the feeling of familiarity. *Memory & Cognition*, *33*(2), 249–260.

Field, A. (2017). *Discovering statistics using IBM SPSS statistics.* Thousand Oaks, CA, Vereinigte Staaten: Sage Publications.

Figl, K. (2010). Deutschsprachige Fragebögen zur Usability-Evaluation im Vergleich. *Zeitschrift für Arbeitswissenschaft*, *4*, 321–337.

Figueroa, P., & Castro, D. (2011). A reusable library of 3D interaction techniques. In *2011 IEEE Symposium on 3D User Interfaces (3DUI)* (S. 3–10). IEEE. Piscataway, NJ, Vereinigte Staaten.

Finsen, L., Søgaard, K., Jensen, C., Borg, V., & Christensen, H. (2001). Muscle activity and cardiovascular response during computer-mouse work with and without memory demands. *Ergonomics*, *44*(14), 1312–1329.

Fischbach, M., Neff, M., Pelzer, I., Lugrin, J.-L., & Latoschik, M. E. (2013). Input device adequacy for multimodal and bimanual object manipulation in virtual environments. In *Virtuelle und Erweiterte Realität, 10. Workshop der GI-Fachgruppe VR/AR.* (S. 145–156). Shaker. Düren, Deutschland.

Fischbein, H. (1987). *Intuition in science and mathematics: An educational approach.* Heidelberg, Deutschland: Springer Science & Business Media.

Fishbach, A., Friedman, R. S., & Kruglanski, A. W. (2003). Leading us not into temptation: Momentary allurements elicit overriding goal activation. *Journal of Personality and Social Psychology*, *84*(2), 296.

Fiske, D. W. (1982). Convergent-discriminant validation in measurements and research strategies. *New Directions for Methodology of Social & Behavioral Science*, *12*, 77–92.

Fisseni, H. (1997). *Lehrbuch der psychologischen Diagnostik.* Göttingen, Deutschland: Hogrefe Verlag.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *51*(4), 327.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906.

Flick, U. (2006). *Qualitative Evaluationsforschung: Konzepte, Methoden, Umsetzungen.* Reinbek bei Hamburg, Deutschland: Rowohlt.

Flick, U. (2010). Gütekriterien qualitativer Forschung. In *Handbuch qualitative Forschung in der Psychologie* (S. 395–407). Heidelberg, Deutschland: Springer.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA, Vereinigte Staaten: MIT Press.

Forster, M., Leder, H., & Ansorge, U. (2013). It felt fluent, and I liked it: Subjective feeling of fluency rather than objective fluency determines liking. *Emotion*, *13*(2), 280.

Forster, M., Leder, H., & Ansorge, U. (2016). Exploring the Subjective Feeling of Fluency. *Experimental Psychology*, *63*(1), 45–58.

Foss, B. M., & Dodwell, P. C. (1966). *New horizons in psychology.* London, Vereinigtes Königreich: Penguin Books.

Fournier, L. R., Wilson, G. F., & Swain, C. R. (1999). Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: Manipulations of task difficulty and training. *International Journal of Psychophysiology*, *31*(2), 129–145.

Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass*, *5*(10), 914–926.

Frese, M., & Zapf, D. (1994). Action as the core of work psychology: A German approach. *Handbook of Industrial and Organizational Psychology*, *4*(2), 271–340.

Friard, O., & Gamba, M. (2016). BORIS: A free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution*, *7*(11), 1325–1330.

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: a latent-variable analysis. *Journal of Experimental Psychology*, *133*(1), 101.

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*, *86*, 186–204.

Fu, L., Salvendy, G., & Turley, L. (2002). Effectiveness of user testing and heuristic evaluation as a function of performance classification. *Behaviour & Information Technology*, *21*(2), 137–143.

Galley, N. (2001). Physiologische Grundlagen, Meßmethoden und Indikatorfunktion der okulomotorischen Aktivität. *Enzyklopädie der Psychologie*, *4*, 237–315.

Galy, E., Cariou, M., & Mélan, C. (2012). What is the relationship between mental workload factors and cognitive load types? *International Journal of Psychophysiology*, *83*(3), 269–275.

Galy, E., Paxion, J., & Berthelon, C. (2018). Measuring mental workload with the NASA-TLX needs to examine each dimension rather than relying on the global score: An example with driving. *Ergonomics*, *61*(4), 517–527.

Gambill, H. D., Ogle, K. N., & Kearns, T. P. (1967). Mydriatic effect of four drugs determined with pupillograph. *Archives of Ophthalmology*, *77*(6), 740–746.

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—in search of a phenomenon. *Thinking & Reasoning*, *21*(4), 383–396.

Gao, Q., Wang, Y., Song, F., Li, Z., & Dong, X. (2013). Mental workload measurement for emergency operating procedures in digital nuclear power plants. *Ergonomics*, *56*(7), 1070–1085.

Garcia, F. A. A. (2012). *Tests to identify outliers in data series* (Unveröffentlichte Dissertation), Pontifical Catholic University of Rio de Janeiro, Industrial Engineering Department, Rio de Janeiro, Brasilien.

Garner, R., Gillingham, M. G., & White, C. S. (1989). Effects of 'seductive details' on macroprocessing and microprocessing in adults and children. *Cognition and Instruction*, *6*(1), 41–57.

Gawron, V. J. (2019). *Human Performance and Situation Awareness Measures*. Boca Raton, FL, Vereinigte Staaten: CRC Press.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692.

Gebhardt, A. (2017). *Additive Fertigungsverfahren: Additive Manufacturing und 3D-Drucken für Prototyping-Tooling-Produktion*. München, Deutschland: Carl Hanser Verlag.

Gediga, G., & Hamborg, K.-C. (2002). Evaluation in der Software-Ergonomie: Methoden und Modelle im Software-Entwicklungsprozess. *Zeitschrift für Psychologie*, *210*(1), 40–57.

Gerrig, R. J., & Zimbardo, P. G. (2008). *Psychologie*. Hallbergmoos, Deutschland: Pearson Studium.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. London, Vereinigtes Königreich: Penguin.

Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and Instruction*, *15*(4), 313–331.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237–288.

Glöckner, A., & Betsch, T. (2008). Multiple-reason decision making based on automatic processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1055.

Glöckner, A., & Witteman, C. (2010). Beyond dual-process models: A categorisation of processes underlying intuitive judgement and decision making. *Thinking & Reasoning*, *16*(1), 1–25.

Goldhammer, F., & Moosbrugger, H. (2006). Aufmerksamkeit. In *Leistung und Leistungsdiagnostik* (S. 16–33). Heidelberg, Deutschland: Springer.

Gooding, P., Isaac, C., & Mayes, A. (2005). Prose recall and amnesia: More implications for the episodic buffer. *Neuropsychologia*, *43*(4), 583–587.

Gopher, D., & Donchin, E. (1986). Workload: An examination of the concept. In *Handbook of Perception and Human Performance, Vol II: Cognitive Processes and Performance*. Hoboken, NJ, Vereinigte Staaten: John Wiley & Sons.

Görner, C., & Ilg, R. (1993). Evaluation der Mensch-Rechner-Schnittstelle. *Benutzergerechte Software-Gestaltung: Standards, Methoden und Werkzeuge*, 189–203.

Graf, L. K., Mayer, S., & Landwehr, J. R. (2018). Measuring processing fluency: One versus five items. *Journal of Consumer Psychology*, *28*(3), 393–411.

Grassmann, M., Vlemincx, E., von Leupoldt, A., Mittelstädt, J. M., & Van den Bergh, O. (2016). Respiratory changes in response to cognitive load: A systematic review. *Neural Plasticity*, *2016*.

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, *13*(3), 203–261.

Gray, W. D., Young, R. M., & Kirschenbaum, S. S. (1997). Introduction to this special issue on Cognitive Architectures and Human-Computer. *Human–Computer Interaction*, *12*(4), 301–309.

Green, E. J., & Barber, P. J. (1981). An auditory Stroop effect with judgments of speaker gender. *Perception & Psychophysics*, *30*(5), 459–466.

Green, E. J., & Barber, P. J. (1983). Interference effects in an auditory Stroop task: Congruence and correspondence. *Acta Psychologica*, *53*(3), 183–194.

Green, M., & Lo, J. (2004). The Grappl 3D interaction technique library. In *Proceedings of the ACM Symposium on Virtual reality Software and Technology* (S. 16–23). ACM. New York, NY, Vereinigte Staaten.

Greenfield, S. (2000). Brain Story: Unlocking our inner world of emotions, memories, ideas and desires. *London, Vereinigtes Königreich: BBC Worldwide*.

Grier, R. A. (2015). How high is high? A meta-analysis of NASA-TLX global workload scores. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Bd. 59, S. 1727–1731). Los Angeles, CA, Vereinigte Staaten: Sage Publications.

Gudur, R. R., Blackler, A., Popovic, V., & Mahar, D. P. (2009). Redundancy in interface design and its impact on intuitive use of a product in older users. In *Proceedings of the 5th IASDR (The International Association of Societies of Design Research Congress), IASDR 2009* (S. 209–209). IASDR. Seoul, Korea.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282.

Gwizdka, J. (2010). Using Stroop task to assess cognitive load. In *Proceedings of the 28th Annual European Conference on Cognitive Ergonomics* (S. 219–222). ACM. New York, NY, Vereinigte Staaten.

Haapalainen, E., Kim, S., Forlizzi, J. F., & Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing* (S. 301–310). ACM. New York, NY, Vereinigte Staaten.

Haar, A. (2013). *Developing of a qualitative classification method for usability errors after Rasmussen.* (Unveröffentlichte Bachelor Thesis), University of Twente, Enschede, Niederlande.

Haase, C. M., Heckhausen, J., & Wrosch, C. (2013). Developmental regulation across the life span: Toward a new synthesis. *Developmental Psychology*, *49*(5), 964.

Hacker, W. (1986). *Arbeitspsychologie.* Berlin, Deutschland: Deutscher Verlag der Wissenschaften.

Hacker, W. (2005). *Allgemeine Arbeitspsychologie: Psychische Regulation von Wissens-, Denk-und körperlicher Arbeit.* Mannheim, Deutschland: Huber.

Hacker, W. (2009). *Arbeitsgegenstand Mensch: Psychologie dialogisch-interaktiver Erwerbsarbeit. Ein Lehrbuch.* Lengerich, Deutschland: PABST Science Publishers.

Hacker, W., & Sachse, P. (2013). *Allgemeine Arbeitspsychologie: Psychische Regulation von Tätigkeiten.* Mannheim, Deutschland: Huber.

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *54*(3), 761–771.

Hadie, S. N., & Yusoff, M. S. (2016). Assessing the validity of the cognitive load scale in a problem-based learning setting. *Journal of Taibah University Medical Sciences*, *11*(3), 194–202.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814.

Hamborg, K.-C., Hoemske, T., & Ollermann, F. (2006). Qualitätssicherung im Usability-Testing-zur Reliabilität eines Klassifikationssystems für Nutzungsprobleme. *Mensch und Computer 2006: Mensch und Computer im Strukturwandel.*

Hammond, K. R. (1993). Naturalistic decision making from a Brunswikian viewpoint: Its past, present, future. *Decision making in action: Models and methods*, 205–227.

Hancock, P. A., & Matthews, G. (2019). Workload and performance: Associations, insensitivities, and dissociations. *Human Factors*, *61*(3), 374–392.

Hand, C. (1997). A survey of 3D interaction techniques. In *Computer Graphics Forum* (Bd. 16, S. 269–281). Hoboken, NJ, Vereinigte Staaten: John Wiley & Sons.

Handley, S. J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: A test of the default interventionist account of belief bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 28.

Hankins, T. C., & Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation, Space, and Environmental Medicine*, *69*(4), 360–367.

Harmon-Jones, E., & Allen, J. J. (2001). The role of affect in the mere exposure effect: Evidence from psychophysiological and individual differences approaches. *Personality and Social Psychology Bulletin*, *27*(7), 889–898.

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology*, *90*(3), 414.

Harris Sr, R. L., Tole, J. R., Ephrath, A. R., & Stephens, A. T. (1982). How a new instrument affects pilots' mental workload. In *Proceedings of the Human Factors Society Annual Meeting* (Bd. 26, *11*, S. 1010–1013). Sage Publications. Los Angeles, CA, Vereinigte Staaten.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, *56*(4), 208.

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Bd. 50, *9*, S. 904–908). Sage Publications. Los Angeles, CA, Vereinigte Staaten.

Hart, S. G., McPherson, D., & Loomis, L. (1978). Time estimation as a secondary task to measure workload: Summary of research. In *Proceedings of the 11th Annual Conference on Manual Control* (S. 64–77). US Government Printing Office. Washington DC, Vereinigte Staaten.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology* (Bd. 52, S. 139–183). Amsterdam, Niederlande: Elsevier.

Hart, S. G., & Wickens, C. D. (1990). Workload assessment and prediction. In *Manprint* (S. 257–296). Heidelberg, Deutschland: Springer.

Hartson, H. R., Andre, T. S., Williges, R. C., & Van Rens, L. (1999). The User Action Framework: A Theory-Based Foundation for Inspection and Classification of Usability Problems. *HCI (1)*, *1999*, 1058–1062.

Hartson, H. R., Andre, T. S., & Williges, R. C. (2001). Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction*, *13*(4), 373–410.

Hartson, H. R., & Castillo, J. C. (1998). Remote evaluation for post-deployment usability improvement. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (S. 22–29). ACM. New York, NY, Vereinigte Staaten.

Hasan, L., Morris, A., & Probets, S. (2012). A comparison of usability evaluation methods for evaluating e-commerce websites. *Behaviour & Information Technology*, *31*(7), 707–737.

Hasselhorn, M., & Artelt, C. (2018). *Metakognition*. Weinheim, Deutschland: Beltz.

Hasselhorn, M., & Grube, D. (2003). Das Arbeitsgedächtnis: Funktionsweise, Entwicklung und Bedeutung für kognitive Leistungsstörungen. *Sprache· Stimme· Gehör*, *27*(01), 31–37.

Hassenzahl, M., & Burmester, M. (1999). Zur Diagnose von Nutzungsproblemen: Praktikable Ansätze aus der qualitativen Forschungspraxis. In *Konferenzband der ZMMS Konferenz* (S. 5–10). Technische Universität Berlin. Berlin, Deutschland.

Hassenzahl, M., & Tractinsky, N. (2006). User experience - A research agenda. *Behaviour & Information Technology*, *25*(2), 91–97.

Havlicek, L. L., & Peterson, N. L. (1976). Robustness of the Pearson correlation against violations of assumptions. *Perceptual and Motor Skills*, *43*(3), 1319–1334.

Heffernan, S. (2013). Video.js (Version 4.3) [Software]. Abgerufen von https://videojs.com/.

Hegarty, M., Shah, P., & Miyake, A. (2000). Constraints on using the dual-task methodology to specify the degree of central executive involvement in cognitive tasks. *Memory & Cognition*, *28*(3), 376–385.

Hegner, M. (2003). Methoden zur Evaluation von Software. *IZ-Arbeitsbericht*, *29*.

Heimbeck, D., Frese, M., Sonnentag, S., & Keith, N. (2003). Integrating errors into the training process: The function of error management instructions and the role of goal orientation. *Personnel Psychology*, *56*(2), 333–361.

Hendy, K. C., Hamilton, K. M., & Landry, L. N. (1993). Measuring subjective workload: When is one scale better than many? *Human Factors*, *35*(4), 579–601.

Hertzum, M., Hansen, K. D., & Andersen, H. H. (2009). Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology*, *28*(2), 165–181.

Hertzum, M., & Jacobsen, N. E. (2001). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, *13*(4), 421–443.

Hertzum, M., Molich, R., & Jacobsen, N. E. (2014). What you get is what you see: Revisiting the evaluator effect in usability tests. *Behaviour & Information Technology*, *33*(2), 144–162.

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner Jr, A. C., Zaklade, A. L., & Christ, R. E. (1992). Comparison of four subjective workload rating scales. *Human Factors*, *34*(4), 429–439.

Hinckley, K., Pausch, R., Goble, J. C., & Kassell, N. F. (1994). A survey of design issues in spatial input. In *Proceedings of the 7th Annual ACM Symposium on User interface Software and Technology* (S. 213–222). New York, NY, Vereinigte Staaten: ACM.

Hjortskov, N., Rissén, D., Blangsted, A. K., Fallentin, N., Lundberg, U., & Søgaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology*, *92*(1-2), 84–89.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology*, *45*(1), 73–93.

Hodgkinson, G. P., Langan-Fox, J., & Sadler-Smith, E. (2008). Intuition: A fundamental bridging construct in the behavioural sciences. *British Journal of Psychology*, *99*(1), 1–27.

Hogarth, R. M. (2001). *Educating intuition*. Chicago, IL, Vereinigte Staaten: University of Chicago Press.

Hohnsbein, J., Falkenstein, M., & Hoormann, J. (1995). Effects of attention and time-pressure on P300 subcomponents and implications for mental workload research. *Biological Psychology*, *40*(1-2), 73–81.

HolidayCheck AG. (2017). HolidayCheck. Abgerufen von https://www.holidaycheck.de/.

Holtzblatt, K., & Beyer, H. (2016). *Contextual design: Design for life.* Burlington, MA, Vereinigte Staaten: Morgan Kaufmann.

Horn, A.-M. (2008). *Validierung des „Questionnaire for Intuitive Use“ am Beispiel der Benutzung zweier Versionen des BVG- Fahrkartenautomaten* (Unveröffentlichte Diplomarbeit), Technische Universität Berlin, Berlin, Deutschland.

Hornbæk, K., & Frøkjær, E. (2008). A study of the evaluator effect in usability testing. *Human–Computer Interaction*, *23*(3), 251–277.

Horrey, W. J., & Wickens, C. D. (2004). Driving and side task performance: The effects of display clutter, separation, and modality. *Human Factors*, *46*(4), 611–624.

Horstmann, N. (2012). *Intuition und Deliberation bei der Entscheidungsfindung: Eine Betrachtung der Prozessebene* (Unveröffentlichte Dissertation), Universität Mannheim, Mannheim, Deutschland.

Howell, D. C. (2009). *Statistical methods for psychology.* Boston, MA, Vereinigte Staaten: Cengage Learning.

Huo, X., Wang, J., & Ghovanloo, M. (2008). A magneto-inductive sensor based wireless tongue-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *16*(5), 497–504.

Hurtienne, J. (2011). *Image schemas and design for intuitive use: Exploring new guidance for user interface design* (Unveröffentlichte Dissertation), Technische Universität Berlin, Berlin, Deutschland.

Hurtienne, J. (2017). How Cognitive Linguistics Inspires HCI: Image Schemas and Image-Schematic Metaphors. *International Journal of Human–Computer Interaction*, *33*(1), 1–20.

Hurtienne, J., & Blessing, L. (2007). Design for Intuitive Use - Testing image schema theory for user interface design. In *DS 42: Proceedings of ICED 2007, the 16th International Conference on Engineering Design.* (S. 829–830). ICED. Paris, Frankreich.

Hurtienne, J., Horn, A.-M., & Langdon, P. (2010). Facets of prior experience and their impact on product usability for older users. In *Designing Inclusive Interactions* (S. 123–132). Heidelberg, Deutschland: Springer.

Hurtienne, J., Klöckner, K., Diefenbach, S., Nass, C., & Maier, A. (2015). Designing with image schemas: Resolving the tension between innovation, inclusion and intuitive use. *Interacting with Computers*, *27*(3), 235–255.

Hurtienne, J., Weber, K., & Blessing, L. (2008). Prior experience and intuitive use: Image schemas in user centred design. In *Designing Inclusive Futures* (S. 107–116). Heidelberg, Deutschland: Springer.

Hußlein, S., Hurtienne, J., Israel, J. H., Mohs, C., Kindsmüller, M. C., Meyer, H. A., … Pohlmeyer, A. (2007). Intuitive Nutzung - nur ein Schlagwort? *In: Design Report*, *11*(7), 26–27.

Hwang, S.-L., Yau, Y.-J., Lin, Y.-T., Chen, J.-H., Huang, T.-H., Yenn, T.-C., & Hsu, C.-C. (2008). Predicting work performance in nuclear power plants. *Safety Science*, *46*(7), 1115–1124.

Hyman, R. (1982). Quasi-experimentation: Design and analysis issues for field settings. *Journal of Personality Assessment*, *46*(1), 96–97.

Iani, C., Gopher, D., Grunwald, A., & Lavie, P. (2007). Peripheral arterial tone as an on-line measure of load in a simulated flight task. *Ergonomics*, *50*(7), 1026–1035.

Iani, C., Gopher, D., & Lavie, P. (2004). Effects of task difficulty and invested mental effort on peripheral vasoconstriction. *Psychophysiology*, *41*(5), 789–798.

Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers.* Milwaukee, WI, Vereinigte Staaten: ASQ Press.

Ikehara, C. S., & Crosby, M. E. (2005). Assessing cognitive load with physiological sensors. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences* (295a–295a). IEEE. Piscataway, NJ, Vereinigte Staaten.

IPO.Plan GmbH. (2015). IPO.Rack (Version 1.1.4) [Software]. Abgerufen von https://www.ipoplan.de/software-ipolog/iporack/.

Ishigaki, H., Miyao, M., & Ishihara, S. (1991). Change of pupil size as a function of exercise. *Journal of Human Ergology*, *20*(1), 61.

ISO. (1998). *DIN EN ISO 9241-11: Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten: Anforderungen an die Gebrauchstauglichkeit–Leitsätze.* Berlin, Deutschland: Beuth Verlag.

ISO. (2011). *DIN EN ISO 9241-210: Ergonomie der Mensch-System-Interaktion-Teil 210: Prozess zur Gestaltung gebrauchstauglicher interaktiver Systeme.* Berlin, Deutschland: Beuth Verlag.

Israel, J. H., Hurtienne, J., Pohlmeyer, A. E., Mohs, C., Kindsmuller, M., & Naumann, A. (2009). On intuitive use, physicality and tangible user interfaces. *International Journal of Arts and Technology*, *2*(4), 348–366.

Jacobsen, N. E., Hertzum, M., & John, B. E. (1998). The evaluator effect in usability studies: Problem detection and severity judgments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Bd. 42, S. 1336–1340). Thousand Oaks, CA, Vereinigte Staaten: Sage Publications.

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology*, *118*(2), 126.

Jahn, G., Oehme, A., Krems, J. F., & Gelau, C. (2005). Peripheral detection as a workload measure in driving: Effects of traffic complexity and route guidance system use in a driving study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*(3), 255–275.

Jankowski, J., & Hachet, M. (2015). Advances in interaction with 3d environments. In *Computer Graphics Forum* (Bd. 34, S. 152–190). London, Vereinigtes Königreich: John Wiley & Sons.

Jeffries, R., & Desurvire, H. (1992). Usability testing vs. heuristic evaluation: Was there a contest? *ACM SIGCHI Bulletin*, *24*(4), 39–41.

Jerger, S., Stout, G., Kent, M., Albritton, E., Loiselle, L., Blondeau, R., & Jorgenson, S. (1993). Auditory Stroop effects in children with hearing impairment. *Journal of Speech, Language, and Hearing Research*, *36*(5), 1083–1096.

Jex, H., McDonnell, J., & Phatak, A. (1966). A "Critical" Tracking Task for Manual Control Research. *IEEE Transactions on Human Factors in Electronics*, (4), 138–145.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge, MA, Vereinigte Staaten: Harvard University Press.

Johnson, M. (2013). *The body in the mind: The bodily basis of meaning, imagination, and reason*. Chicago, IL, Vereinigte Staaten: University of Chicago Press.

Jonassen, D. H., & Grabowski, B. L. H. (1993). *Handbook of Individual Differences, Learning, and Instruction*. Abingdon, Vereinigtes Königreich: Routledge.

Jurado, M. B., & Rosselli, M. (2007). The elusive nature of executive functions: A review of our current understanding. *Neuropsychology Review*, *17*(3), 213–233.

Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *47*(2), 310.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ, Vereinigte Staaten: Prentice-Hall.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697.

Kahneman, D., & Egan, P. (2011). *Thinking, fast and slow*. New York, NY, Vereinigte Staaten: Macmillan Publishers.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and Biases: The Psychology of Intuitive Judgment*, *49*, 81.

Kanis, H. (2014). Reliability and validity of findings in ergonomics research. *Theoretical Issues in Ergonomics Science*, *15*(1), 1–46.

Kaptein, M., & Robertson, J. (2012). Rethinking statistical analysis methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (S. 1105–1114). ACM. New York, NY, Vereinigte Staaten.

Kapur, A., Kapur, S., & Maes, P. (2018). Alterego: A personalized wearable silent speech interface. In *23rd International Conference on Intelligent User Interfaces* (S. 43–53). New York, NY, Vereinigte Staaten: ACM.

Karar, V., & Ghosh, S. (2014). Attention tunneling: Effects of limiting field of view due to beam combiner frame of head-up display. *Journal of Display Technology*, *10*(7), 582–589.

Kauffeld, S. (2014). Einführung in die Arbeits-, Organisations-und Personalpsychologie. In *Arbeits-, Organisations-und Personalpsychologie für Bachelor* (S. 1–14). Heidelberg, Deutschland: Springer.

Keele, S. W. (1973). *Attention and human performance*. Pacific Palisades, CA, Vereinigte Staaten: Goodyear Publishing Company.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, *4*(6), 533–550.

Kerkau, F. (2006). *Biosignale der Pupille zur Steuerung intelligenter User-Interfaces: Untersuchung von Pupillenbewegungen zur Realisierung einer biopsychologischen Computerschnittstelle für die Mensch Computer Interaktion* (Unveröffentlichte Dissertation), Freie Universität Berlin, Berlin, Deutschland.

Kerr, B. (1973). Processing demands during mental operations. *Memory & Cognition*, *1*(4), 401–412.

Kerr, B., Condon, S. M., & McDonald, L. A. (1985). Cognitive spatial processing and the regulation of posture. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(5), 617.

Kiefer, M., Marzinzik, F., Weisbrod, M., Scherg, M., & Spitzer, M. (1998). The time course of brain activations during response inhibition: Evidence from event-related potentials in a go/no go task. *Neuroreport*, *9*(4), 765–770.

Klaczynski, P. A. (2000). Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: A two-process approach to adolescent cognition. *Child Development*, *71*(5), 1347–1366.

Klein, G. (1993). *A recognition-primed decision (RPD) model of rapid decision making.* New York, NY, Vereinigte Staaten: Ablex Publishing Corporation.

Klein, G. (1998). *Sources of Power: How people make decisions.* Cambridge, MA, Vereinigte Staaten: MIT Press.

Klein, G. (2008). Naturalistic decision making. *Human Factors*, *50*(3), 456–460.

Klimesch, W. (1997). EEG-alpha rhythms and memory processes. *International Journal of Psychophysiology*, *26*(1-3), 319–340.

Kobus, D. A., Russotti, J., Schlichting, C., Haskell, G., Carpenter, S., & Wojtowicz, J. (1986). Multimodal detection and recognition performance of sonar operators. *Human Factors*, *28*(1), 23–29.

Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology*, *38*(3), 557–577.

König, C. (2012). *Analyse und Anwendung eines menschzentrierten Gestaltungsprozesses zur Entwicklung von Human-Machine-Interfaces im Arbeitskontext am Beispiel Flugsicherung* (Unveröffentlichte Dissertation), Technische Universität Darmstadt, Darmstadt, Deutschland.

Kopp, B., & Mandl, H. (2005). *Wissensschemata* (Techn. Ber. Nr. 177). LMU München. München, Deutschland.

Korbach, A., Brünken, R., & Park, B. (2018). Differentiating different types of cognitive load: A comparison of different measures. *Educational Psychology Review*, *30*(2), 503–529.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*(4), 609.

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, *9*(2), 149–171.

Kosch, T., Hassib, M., Buschek, D., & Schmidt, A. (2018). Look into my eyes: using pupil dilation to estimate mental workload for task complexity adaptation. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (S. 1–6). ACM. New York, NY, Vereinigte Staaten.

Koutsabasis, P., Spyrou, T., & Darzentas, J. (2007). Evaluating usability evaluation methods: Criteria, method and a case study. In *International Conference on Human-Computer Interaction* (S. 569–578). Heidelberg, Deutschland: Springer.

Kraft, J. F., & Hurtienne, J. (2017). Transition animations support orientation in mobile interfaces without increased user effort. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (S. 17). ACM. New York, NY, Vereinigte Staaten.

Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. *Multiple-Task Performance*, 279–328.

Kramer, A. F., & Spinks, J. (1991). Capacity views of human information processing. *Handbook of Cognitive Psychophysiology, Central and Autonomic Nervous System Approaches.* 179–249.

Krause, U.-M., & Stark, R. (2006). Vorwissen aktivieren. In *Handbuch Lernstrategien* (S. 38–49). Göttingen, Deutschland: Hogrefe Verlag.

Krohne, H. W., & Hock, M. (2007). *Psychologische Diagnostik: Grundlagen und Anwendungsfelder*. Stuttgart, Deutschland: Kohlhammer Verlag.

Kruglanski, A. W. (2013). Only one? The default interventionist perspective as a unimodel — Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, *8*(3), 242–247.

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review*, *118*(1), 97.

Kuhn, T. L. (1974). Discrimination of modulated beat tempo by professional musicians. *Journal of Research in Music Education*, *22*(4), 270–277.

Kunert, R., Willems, R. M., & Hagoort, P. (2016). An independent psychometric evaluation of the PROMS measure of music perception skills. *PloS One*, *11*(7), e0159103.

Kuusela, H., & Pallab, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *The American Journal of Psychology*, *113*(3), 387.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Lanzilotti, R., Ardito, C., Costabile, M. F., & De Angeli, A. (2011). Do patterns help novice evaluators? A comparative study. *International Journal of Human-Computer Studies*, *69*(1), 52–69.

LaViola, J. J., Kruijff, E., McMahan, R. P., Bowman, D., & Poupyrev, I. P. (2017). *3D user interfaces: Theory and practice*. Boston, MA, Vereinigte Staaten: Addison-Wesley.

Law, L. N., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the Profile of Music Perception Skills. *PloS One*, *7*(12), e52508.

Lawry, S. (2012). *Identifying familiarity to facilitate intuitive interaction for older adults* (Unveröffentlichte Dissertation), Queensland University of Technology, Brisbane, Australien.

Lawry, S., Popovic, V., & Blackler, A. L. (2011). Diversity in product familiarity across younger and older adults. In *Diversity & Unity: Proceedings of the 4th IASDR (The International Association of Societies of Design Research Congress), IASDR2011*. IASDR, Delft, Niederlande.

Lawry, S., Popovic, V., Blackler, A., & Thompson, H. (2019). Age, familiarity, and intuitive use: An empirical investigation. *Applied Ergonomics*, *74*, 74–84.

Lee, J. D., Wickens, C. D., Liu, Y., & Boyle, L. N. (2017). *Designing for people: An introduction to human factors engineering*. Scotts Valley, CA, Vereinigte Staaten: CreateSpace.

Lee, Y.-H., & Liu, B.-S. (2003). Inflight workload assessment: Comparison of subjective and physiological measurements. *Aviation, Space, and Environmental Medicine*, *74*(10), 1078–1084.

Legewie, H. (1994). *Globalauswertung von Dokumenten*. Konstanz, Deutschland: UVK Verlagsgesellschaft.

Lehman, S., Schraw, G., McCrudden, M. T., & Hartley, K. (2007). Processing and recall of seductive details in scientific text. *Contemporary Educational Psychology*, *32*(4), 569–587.

Lewis, J. R. (2014). Usability: lessons learned... and yet to be learned. *International Journal of Human-Computer Interaction*, *30*(9), 663–684.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348.

Lieberman, M. D. (2000). Intuition: A social cognitive neuroscience approach. *Psychological Bulletin*, *126*(1), 109.

Lieberman, M. D. (2003). Reflective and reflexive judgment processes: A social cognitive neuroscience approach. In J. Forgas, K. Williams, & W. Von Hippel (Hrsg.), *Social judgments: Implicit and explicit processes* (S. 44–67). Cambridge, Vereinigtes Königreich: Cambridge University Press.

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse.* Weinheim, Deutschland: Beltz.

Lim, K. H., Ward, L. M., & Benbasat, I. (1997). An empirical study of computer system learning: Comparison of co-discovery and self-discovery methods. *Information Systems Research*, *8*(3), 254–272.

LimeSurvey Development Team. (2015). LimeSurvey (Version 2.05) [Software]. Abgerufen von https://www.limesurvey.org/.

Lindgaard, G. (2006). Notions of thoroughness, efficiency, and validity: Are they valid in HCI practice? *International Journal of Industrial Ergonomics*, *36*(12), 1069–1074.

Lipp, O. V., & Neumann, D. L. (2004). Attentional blink reflex modulation in a continuous performance task is modality specific. *Psychophysiology*, *41*(3), 417–425.

Liu, Y., & Wickens, C. D. (1992). Visual scanning with or without spatial uncertainty and divided and selective attention. *Acta Psychologica*, *79*(2), 131–153.

Liu, Y., & Wickens, C. D. (1994). Mental workload and cognitive task automaticity: An evaluation of subjective and time estimation metrics. *Ergonomics*, *37*(11), 1843–1854.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research*, *66*(4), 579–619.

Lodge, M. (1981). *Magnitude Scaling.* Thousand Oaks, CA, Vereinigte Staaten: Sage Publications.

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, *65*(3), 272–292.

Löffler, D., Hess, A., Maier, A., Hurtienne, J., & Schmitt, H. (2013). Developing Intuitive User Interfaces by Integrating Users' Mental Models into Requirements Engineering. In *Proceedings of the 27th International BCS Human Computer Interaction Conference* (15:1–15:10). BCS-HCI '13. Swinton, Vereinigtes Königreich: British Computer Society.

Logie, R. H. (2011). The functional organization and capacity limits of working memory. *Current Directions in Psychological Science*, *20*(4), 240–245.

Logie, R. H. (2014). *Visuo-spatial working memory.* London, Vereinigtes Königreich: Psychology Press.

Longo, L. (2014). *Formalising human mental workload as a defeasible computational concept* (Unveröffentlichte Dissertation), Dublin Institute of Technology, Dublin, Irland.

Longo, L. (2015). A defeasible reasoning framework for human mental workload representation and assessment. *Behaviour & Information Technology*, *34*(8), 758–786.

Longo, L. (2017). Subjective usability, mental workload assessments and their impact on objective human performance. In *IFIP Conference on Human-Computer Interaction* (S. 202–223). Springer. Heidelberg, Deutschland.

Longo, L., & Dondio, P. (2015). On the relationship between perception of usability and subjective mental workload of web interfaces. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on* (Bd. 1, S. 345–352). Piscataway, NJ, Vereinigte Staaten: IEEE.

Longo, L., & Leva, M. C. (2017). *Human Mental Workload: Models and Applications.* Heidelberg, Deutschland: Springer.

Lund, N. (2001). *Attention and Pattern Recognition.* London, Vereinigtes Königreich: Psychology Press.

Lysaght, R. J., Hill, S. G., Dick, A., Plamondon, B. D., & Linton, P. M. (1989). *Operator workload: Comprehensive review and evaluation of operator workload methodologies.* Willow Grove, PA, Vereinigte Staaten: Analytics Inc.

Macaranas, A. (2013). *The effects of intuitive interaction mappings on the usability of body-based interfaces* (Unveröffentlichte Dissertation), Queensland University of Technology, Brisbane, Australien.

Macaranas, A., Antle, A. N., & Riecke, B. E. (2015). What is intuitive interaction? Balancing users' performance and satisfaction with natural user interfaces. *Interacting with Computers*, *27*(3), 357–370.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163–203.

Madden, K., & Savard, G. (1995). Effects of mental state on heart rate and blood pressure variability in men and women. *Clinical Physiology*, *15*(6), 557–569.

Madsen, C. K. (1979). Modulated beat discrimination among musicians and nonmusicians. *Journal of Research in Music Education*, *27*(2), 57–67.

Makri, S., Blandford, A., Cox, A. L., Attfield, S., & Warwick, C. (2011). Evaluating the Information Behaviour methods: Formative evaluations of two methods for assessing the functionality and usability of electronic information resources. *International Journal of Human-Computer Studies*, *69*(7), 455–482.

Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review*, *99*(4), 587.

Mandler, J. M. (2005). How to build a baby: III. Image schemas and the transition to verbal thought. *From Perception to Meaning: Image schemas in Cognitive Linguistics*, 137–163.

Mandler, J. M. (2014). *Stories, scripts, and scenes: Aspects of schema theory.* London, Vereinigtes Königreich: Psychology Press.

Mangan, B. (2015). The uncanny valley as fringe experience. *Interaction Studies*, *16*(2), 193–199.

Manzey, D. (1997). Psychologie mentaler Beanspruchung. In F. Rösler (Hrsg.), *Ergebnisse und Anwendungen der Psychophysiologie. Enzyklopädie der Psychologie* (S. 799–864). Göttingen, Deutschland: Hogrefe.

Markovits, H., Thompson, V. A., & Brisson, J. (2015). Metacognition and abstract reasoning. *Memory & Cognition*, *43*(4), 681–693.

Marshall, P., Rogers, Y., & Hornecker, E. (2007). Are tangible interfaces really any better than other kinds of interfaces? In *Extended CHI'07 Abstracts on Human Factors in Computing Systems* (S. 2817–2820). ACM. New York, NY, Vereinigte Staaten.

Marshall, S. P. (2002). The index of cognitive activity: Measuring cognitive workload. In *Proceedings of the IEEE 7th Conference on Human Factors and Power Plants* (S. 7–7). IEEE. Piscataway, NJ, Vereinigte Staaten.

Marshall, S. P., Pleydell-Pearce, C. W., & Dickson, B. T. (2003). Integrating psychophysiological measures of cognitive workload and eye movements to detect strategy shifts. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the* (6–pp). IEEE. Piscataway, NJ, Vereinigte Staaten.

Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Abich IV, J. (2015). The psychometrics of mental workload: Multiple measures are sensitive but divergent. *Human Factors*, *57*(1), 125–143.

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs? *Memory & Cognition*, *29*(2), 222–233.

May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., & Brannan, J. R. (1990). Eye movement indices of mental workload. *Acta Psychologica*, *75*(1), 75–89.

Mayo, C. W., & Crockett, W. H. (1964). Cognitive complexity and primacy-recency effects in impression formation. *The Journal of Abnormal and Social Psychology*, *68*(3), 335.

Mazzoni, D. (2014). Audacity (Version 1.3.5) [Software]. Abgerufen von https://www.audacity.de/.

McAran, D. (2018). Development of the technology acceptance intuitive interaction model. In *Intuitive Interaction* (S. 129–149). Boca Raton, FL, Vereinigte Staaten: CRC Press.

McClain, L. (1983). Stimulus-response compatibility affects auditory Stroop interference. *Perception & Psychophysics*, *33*(3), 266–270.

McEwan, M. W. (2017). *The influence of naturally mapped control interfaces for video games on the player experience and intuitive interaction* (Unveröffentlichte Dissertation), Queensland University of Technology, Brisbane, Australien.

McEwan, M. W., Blackler, A., Johnson, D. M., & Wyeth, P. A. (2014). Natural mapping and intuitive interaction in videogames. In *Proceedings of the first ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play* (S. 191–200). ACM. New York, NY, Vereinigte Staaten.

McFarlane, D. C., & Latorella, K. A. (2002). The scope and importance of human interruption in human-computer interaction design. *Human-Computer Interaction*, *17*(1), 1–61.

McIsaac, T. L., Lamberg, E. M., & Muratori, L. M. (2015). Building a framework for a dual task taxonomy. *BioMed Research International*, *2015*, 1–10.

McKendrick, R. D., & Cherry, E. (2018). A Deeper Look at the NASA TLX and Where It Falls Short. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Bd. 62, S. 44–48). Los Angeles, CA, Vereinigte Staaten: Sage Publications.

McLeod, P. (1977). A dual task response modality effect: Support for multiprocessor models of attention. *The Quarterly Journal of Experimental Psychology*, *29*(4), 651–667.

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220.

Mega, L. F., Gigerenzer, G., & Volz, K. G. (2015). Do intuitive and deliberate judgments rely on two distinct neural systems? A case study in face processing. *Frontiers in Human Neuroscience*, *9*(25), 456.

Mehler, B., Reimer, B., Coughlin, J., & Dusek, J. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record: Journal of the Transportation Research Board*, (2138), 6–12.

Meijman, T. F. (1991). *Over vermoeidheid: Arbeidspsychologische studies naar de beleving van belastingseffecten* (Unveröffentlichte Dissertation). Amsterdam, Niederlande.

Meshkati, N., & Hancock, P. A. (2011). *Human mental workload.* Amsterdam, Niederlande: Elsevier.

Metcalfe, J. (2000). Feelings and judgments of knowing: Is there a special noetic state? Cambridge, MA, Vereinigte Staaten: Academic Press.

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*, *106*(1), 3–19.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 851–861.

Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & Cognition*, *15*(3), 238–246.

Meyberg, S., Werkle-Bergner, M., Sommer, W., & Dimigen, O. (2015). Microsaccade-related brain potentials signal the focus of visuospatial attention. *NeuroImage*, *104*, 79–88.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81.

Miller, G. A., Galanter, E., & Pribram, K. H. (1973). *Strategien des Handelns: Pläne und Strukturen des Verhaltens.* Stuttgart, Deutschland: Klett.

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception.* Cambridge, MA, Vereinigte Staaten: Belknap Press.

Miller, M. W., Rietschel, J. C., McDonald, C. G., & Hatfield, B. D. (2011). A novel approach to the physiological measurement of mental workload. *International Journal of Psychophysiology*, *80*(1), 75–78.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

Miyake, A., & Shah, P. (1999). *Models of working memory: Mechanisms of active maintenance and executive control.* Cambridge, Vereinigtes Königreich: Cambridge University Press.

Miyake, S., Yamada, S., Shoji, T., Takae, Y., Kuge, N., & Yamamura, T. (2009). Physiological responses to workload change. A test/retest examination. *Applied Ergonomics*, *40*(6), 987–996.

Miyake, Y., Onishi, Y., & Poppel, E. (2004). Two types of anticipation in synchronization tapping. *Acta Neurobiologiae Experimentalis*, *64*(3), 415–426.

Moder, K. (2007). How to keep the Type I Error Rate in ANOVA if Variances are Heteroscedastic. *Austrian Journal of Statistics*, *36*(3), 179–188.

Moder, K. (2010). Alternatives to F-test in one way ANOVA in case of heterogeneity of variances (a simulation study). *Psychological Test and Assessment Modeling*, *52*(4), 343–353.

Möhring, W., & Schlütz, D. (2013). *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft*. Heidelberg, Deutschland: Springer.

Mohs, C., Hurtienne, J., Kindsmüller, M. C., Israel, J. H., Meyer, H. A. et al. (2006). IUUI–Intuitive Use of User Interfaces: Auf dem Weg zu einer wissenschaftlichen Basis für das Schlagwort „Intuitivität". *MMI-Interaktiv*, *11*(11), 75–84.

Mohs, C., Hurtienne, J., Scholz, D., & Rotting, M. (2006). Intuitivitat: Definierbar, beeinflussbar, überprüfbar. *VDI BERICHTE*, *1946*, 215.

Molich, R., & Dumas, J. S. (2008). Comparative usability evaluation (CUE-4). *Behaviour & Information Technology*, *27*(3), 263–281.

Molich, R., Ede, M. R., Kaasgaard, K., & Karyukin, B. (2004). Comparative usability evaluation. *Behaviour & Information Technology*, *23*(1), 65–74.

Molich, R., Thomsen, A. D., Karyukina, B., Schmidt, L., Ede, M., Van Oel, W., & Arcuri, M. (1999). Comparative evaluation of usability tests. In *CHI'99 Extended Abstracts on Human Factors in Computing Systems* (S. 83–84). New York, NY, Vereinigte Staaten: ACM.

Montori, V. M., Swiontkowski, M. F., & Cook, D. J. (2003). Methodologic issues in systematic reviews and meta-analyses. *Clinical Orthopaedics and Related Research*, *413*, 43–54.

Moosbrugger, H., & Kelava, A. (2007). *Testtheorie und Fragebogenkonstruktion*. Heidelberg, Deutschland: Springer.

Moosbrugger, H., & Schermelleh-Engel, K. (2012). Exploratorische (EFA) und konfirmatorische Faktorenanalyse (CFA). In *Testtheorie und Fragebogenkonstruktion* (S. 325–343). Heidelberg, Deutschland: Springer.

Moreno, R. (2010). Cognitive load theory: More food for thought. *Instructional Science*, *38*(2), 135–141.

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in Cognitive Sciences*, *14*(10), 435–440.

Morgan, A. L., & Brandt, J. F. (1989). An auditory Stroop effect for pitch, loudness, and time. *Brain and Language*, *36*(4), 592–603.

Moustafa, K., Luz, S., & Longo, L. (2017). Assessment of mental workload: A comparison of machine learning methods and subjective assessment techniques. In *International Symposium on Human Mental Workload: Models and Applications* (S. 30–50). Springer. Heidelberg, Deutschland.

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review*, *20*(2), 378–384.

Mulder, G., Mulder, L. J., Meijman, T. F., Veldman, J. B., & Van Roon, A. M. (2000). A psychophysiological approach to working conditions. *Engineering Psychophysiology: Issues and Applications*, 139–159.

Mulder, L. J. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology*, *34*, 205–236.

Mulder, L. J., De Waard, D., & Brookhuis, K. A. (2004). Estimating mental effort using heart rate and heart rate variability. In *Handbook of Human Factors and Ergonomics Methods* (S. 227–236). Boca Raton, FL, Vereinigte Staaten: CRC Press.

Murdock Jr, B. B. (1965). Effects of a subsidiary task on short-term memory. *British Journal of Psychology*, *56*(4), 413–419.

Myers, G. J., Sandler, C., & Badgett, T. (2011). *The art of software testing*. Hoboken, NJ, Vereinigte Staaten: John Wiley & Sons.

Naismith, L. M., & Cavalcanti, R. B. (2015). Validity of cognitive load measures in simulation-based training: A systematic review. *Academic Medicine*, *90*(11), S24–S35.

Naumann, A., & Hurtienne, J. (2010). Benchmarks for Intuitive Interaction with Mobile Devices. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services* (S. 401–402). ACM. New York, NY, Vereinigte Staaten.

Naumann, A., Hurtienne, J., Israel, J. H., Mohs, C., Kindsmüller, M. C., Meyer, H., & Hußlein, S. (2007). Intuitive use of user interfaces: Defining a vague concept. In *International Conference on Engineering Psychology and Cognitive Ergonomics* (S. 128–136). Springer. Heidelberg, Deutschland.

Naumann, A., Pohlmeyer, A. E., Husslein, S., Kindsmüller, M. C., Mohs, C., & Israel, J. H. (2008). Design for intuitive use: Beyond usability. In *CHI'08 Extended Abstracts on Human Factors in Computing Systems* (S. 2375–2378). ACM. New York, NY, Vereinigte Staaten.

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review*, *86*(3), 214.

Nee, D. E., Brown, J. W., Askren, M. K., Berman, M. G., Demiralp, E., Krawitz, A., & Jonides, J. (2012). A meta-analysis of executive components of working memory. *Cerebral Cortex*, *23*(2), 264–282.

Nelson, T. O., Leonesio, J., Shimamura, A. P., Landwehr, R. F., & Narens, L. (1982). Overlearning and the feeling of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(4), 279.

Nerdinger, F. W., Blickle, G., Schaper, N., & Schaper, N. (2014). *Arbeits-und Organisationspsychologie*. Heidelberg, Deutschland: Springer.

Netzer, R. H., & Miller, B. P. (1992). What are race conditions?: Some issues and formalizations. *ACM Letters on Programming Languages and Systems (LOPLAS)*, *1*(1), 74–88.

Neumann, O. (1996). Theories of attention. In *Handbook of Perception and Action* (Bd. 3, S. 389–446). Amsterdam, Niederlande: Elsevier.

Newell, A., Simon, H. A. et al. (1972). *Human problem solving*. Englewood Cliffs, NJ, Vereinigte Staaten: Prentice-Hall.

Newell, B. R., Lagnado, D. A., & Shanks, D. R. (2015). *Straight choices: The psychology of decision making*. London, Vereinigtes Königreich: Psychology Press.

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, *37*(1), 1–19.

Newstead, S. E. (2000). Are there two different types of thinking? *Behavioral and Brain Sciences*, *23*(5), 690–691.

Nickel, P., & Nachreiner, F. (2003). Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload. *Human Factors*, *45*(4), 575–590.

Nielsen, J. (1994). *Usability engineering*. Amsterdam, Niederlande: Elsevier.

Nielsen, J. (2006). Quantitative studies: How many users to test. Abgerufen von https://www.nngroup.com/articles/quantitative-studies-how-many-users/.

Nielsen, J. (2012). How Many Test Users in a Usability Study? Abgerufen von https://www.nngroup.com/articles/how-many-test-users/.

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems* (S. 206–213). ACM. New York, NY, Vereinigte Staaten.

Nilsson, M., Drugge, M., Liljedahl, U., Synnes, K., & Parnes, P. (2005). A study on users' preference on interruption when using wearable computers and head mounted displays. In *Third IEEE International Conference on Pervasive Computing and Communications* (S. 149–158). IEEE. Piscataway, NJ, Vereinigte Staaten.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231.

Nittono, H., Hamada, A., & Hori, T. (2003). Brain potentials after clicking a mouse: A new psychophysiological approach to human-computer interaction. *Human Factors*, *45*(4), 591–600.

Noble, M., Fitts, P. M., & Warren, C. E. (1955). The frequency response of skilled subjects in a pursuit tracking task. *Journal of Experimental Psychology*, *49*(4), 249.

Nørgaard, M., & Hornbæk, K. (2006). What do usability evaluators do in practice? An explorative study of think-aloud testing. In *Proceedings of the 6th conference on Designing Interactive Systems* (S. 209–218). ACM. New York, NY, Vereinigte Staaten.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, *7*(1), 44–64.

Norman, D. A., & Shallice, T. (1986). Attention to action. In *Consciousness and Self-regulation* (S. 1–18). Heidelberg, Deutschland: Springer.

Norman, E. (2017). Metacognition and mindfulness: The role of fringe consciousness. *Mindfulness*, *8*(1), 95–100.

Norman, E., Price, M. C., & Duff, S. C. (2010). Fringe consciousness: A useful framework for clarifying the nature of experience-based metacognitive feelings. In *Trends and Prospects in Metacognition Research* (S. 63–80). Heidelberg, Deutschland: Springer.

Nourbakhsh, N., Wang, Y., & Chen, F. (2013). GSR and blink features for cognitive load classification. In *IFIP Conference on Human-Computer Interaction* (S. 159–166). Springer. Heidelberg, Deutschland.

Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. A. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Australian Computer-Human Interaction Conference* (S. 420–423). ACM. New York, NY, Vereinigte Staaten.

O'Brien, M. A., Rogers, W. A., & Fisk, A. D. (2008). Developing a framework for intuitive human-computer interaction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Bd. 52, S. 1645–1649). Los Angeles, CA, Vereinigte Staaten: Sage Publications.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In *Handbook of perception and human performance, Vol. 2: Cognitive processes and performance.* (S. 1–49). Oxford, Vereinigtes Königreich: John Wiley & Sons.

O'Brien, M. A. (2018). Lessons on intuitive usage from everyday technology interactions among younger and older people. In *Intuitive Interaction* (S. 89–111). Boca Raton, FL, Vereinigte Staaten: CRC Press.

O'Brien, M. A., Rogers, W. A., & Fisk, A. D. (2010). *Developing an organizational model for intuitive design* (Techn. Ber. Nr. HFA-TR-1001). Georgia Institute of Technology. Atlanta, GA, Vereinigte Staaten.

Objective Development Software GmbH. (2012). EasyLogger: Sample Application demonstrating how to run V-USB without a crystal, leaving more pins for I/O [Software]. Abgerufen von https://www.obdev.at/products/vusb/easylogger.html.

Ogden, G. D., Levine, J. M., & Eisner, E. J. (1979). Measurement of workload by secondary tasks. *Human Factors*, *21*(5), 529–548.

Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika*, *40*(2), 129–147.

Okimoto, M. L. L., Silva, C. M. A., & Miranda, C. (2012). Approaches to the Intuitive Use in Emergency Situations. *Advances in Usability Evaluation*, 316.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5*(3), 343.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, *11*(6), 988–1010.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429.

Paas, F. G., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63–71.

Paas, F. G., & Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, *6*(4), 351–371.

Pagano, R. (2006). *Understanding statistics in the behavioral sciences.* Boston, MA, Vereinigte Staaten: Cengage Learning.

Palmer, J. M. (2018). Intuitive Interaction in Industry User Research: Context Is Everything. In *Intuitive Interaction* (S. 213–226). Boca Raton, FL, Vereinigte Staaten: CRC Press.

Panach, J. I., Condori-Fernández, N., Valverde, F., Aquino, N., & Pastor, Ó. (2008). Understandability measurement in an early usability evaluation for model-driven development: An empirical study. In *Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement* (S. 354–356). ACM. New York, NY, Vereinigte Staaten.

Pankok Jr, C., Zahabi, M., Zhang, W., Choi, I., Liao, Y.-F., Nam, C. S., & Kaber, D. (2017). The effects of interruption similarity and complexity on performance in a simulated visual-manual assembly operation. *Applied Ergonomics*, *59*, 94–103.

Park, B., & Brünken, R. (2015). The Rhythm Method: A New Method for Measuring Cognitive Load—An Experimental Dual-Task Study. *Applied Cognitive Psychology*, *29*(2), 232–243.

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*(2), 220.

Patterson, R. E. (2017). Intuitive cognition and models of human–automation interaction. *Human Factors*, *59*(1), 101–115.

Paxion, J., Galy, E., & Berthelon, C. (2014). Mental workload and driving. *Frontiers in Psychology*, *5*, 1344.

Penaz, J. (1973). Photoelectric measurement of blood pressure, volume and flow in the finger. In *Digest of the 10th Int. Conf. Med. Biol. Engineering, 1973* (Bd. 104), Dresden, Deutschland.

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 544.

Pfendler, C. (1990). Zur Messung der mentalen Beanspruchung mit dem NASA-Task Load Index. *Zeitschrift für Arbeitswissenschaft*, *44*(3), 158–163.

Pfendler, C. (1991). *Vergleichende Bewertung der NASA-TLX Skala und der ZEIS-Skala bei der Erfassung von Lernprozessen.* Wachtberg, Deutschland: Forschungsinstitut für Anthropotechnik.

Pfister, H.-R., Jungermann, H., & Fischer, K. (2016). *Die Psychologie der Entscheidung: Eine Einführung.* Heidelberg, Deutschland: Springer.

Pfleging, B., Fekety, D. K., Schmidt, A., & Kun, A. L. (2016). A model relating pupil diameter to mental workload and lighting conditions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (S. 5776–5788). ACM. New York, NY, Vereinigte Staaten.

Piaget, J. (1976). *Die Äquilibration der kognitiven Strukturen.* Stuttgart, Deutschland: Klett.

Plessner, H., Betsch, C., & Betsch, T. (2011). *Intuition in judgment and decision making.* London, Vereinigtes Königreich: Psychology Press.

Plessner, H., & Czenna, S. (2011). The benefits of intuition. In *Intuition in judgment and decision making* (S. 271–286). London, Vereinigtes Königreich: Psychology Press.

Poikonen, H., Alluri, V., Brattico, E., Lartillot, O., Tervaniemi, M., & Huotilainen, M. (2016). Event-related brain responses while listening to entire pieces of music. *Neuroscience*, *312*, 58–73.

Pollock, J. L. (1996). OSCAR: A general-purpose defeasible reasoner. *Journal of Applied Non-Classical Logics*, *6*(1), 89–113.

Posner, M. I., & Snyder, C. R. (2004). Attention and cognitive control. *Cognitive Psychology*, *205*.

Preim, B., & Dachselt, R. (2015). *Interaktive Systeme: Band 2: User Interface Engineering, 3D-Interaktion, Natural User Interfaces.* Heidelberg, Deutschland: Springer.

Price, M. C., & Norman, E. (2008). Intuitive decisions on the fringes of consciousness: Are they conscious and does it matter? *Judgment and Decision Making*, *3*(1), 28.

Quinn, J., & McConnell, J. (1996). Irrelevant pictures in visual working memory. *The Quarterly Journal of Experimental Psychology Section A*, *49*(1), 200–215.

Ramacci, C., & Rota, P. (1975). Flight fitness and psycho-physiological behavior of applicant pilots in the first flight missions. *AGARD Med. Requirements and Exam. Procedures in Relation to the Tasks of Today's Aircrew*, *8*, 15–51.

Raskin, J. (1994). Intuitive equals familiar. *Communications of the ACM*, *37*(9), 17–19.

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*, (3), 257–266.

Rasmussen, J. (1986). *Information processing and human-machine interaction.* New York, NY, Vereinigte Staaten: North Holland.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology*, *118*(3), 219.

Reber, R., Fazendeiro, T., & Winkielman, P. (2002). Processing fluency as the source of experiences at the fringe of consciousness. *Psyche*, *8*(10), 1–21.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*(3), 338–342.

Reber, R., & Schwarz, N. (2001). The hot fringes of consciousness: Perceptual fluency and affect. *Consciousness & Emotion*, *2*(2), 223–231.

Reber, R., Wurtz, P., & Zimmermann, T. D. (2004). Exploring "fringe" consciousness: The subjective experience of perceptual fluency and its objective bases. *Consciousness and Cognition*, *13*(1), 47–60.

Recarte, M. Á., Pérez, E., Conchillo, Á., & Nunes, L. M. (2008). Mental workload and visual impairment: Differences between pupil, blink, and subjective rating. *The Spanish Journal of Psychology*, *11*(2), 374–385.

Reddy, G. R. (2012). *Approaches to designing for older adults' intuitive interaction with complex devices* (Unveröffentlichte Dissertation), Queensland University of Technology, Brisbane, Australien.

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In *Advances in Psychology* (Bd. 52, S. 185–218). Amsterdam, Niederlande: Elsevier.

Reimann, M., Zaichkowsky, J., Neuhaus, C., Bender, T., & Weber, B. (2010). Aesthetic package design: A behavioral, neural, and psychological investigation. *Journal of Consumer Psychology*, *20*(4), 431–441.

Reinhardt, D. (2019). GitLab-Repository für das USB-Fußpedal Taktschuh. Abgerufen von https://gitlab2.informatik.uni-wuerzburg.de/s329728/taktschuh.

Reinhardt, D., & Hurtienne, J. (2017). Interaction under pressure: Increased mental workload makes issues of intuitive interaction visible. In *Proceedings of the 2017 ACM Conference Companion Publication on Designing Interactive Systems* (S. 67–71). New York, NY, Vereinigte Staaten: ACM.

Reinhardt, D., & Hurtienne, J. (2018). Cursor Entropy Reveals Decision Fatigue. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion* (S. 31). New York, NY, Vereinigte Staaten: ACM.

Reinhardt, D., Kuge, J., & Hurtienne, J. (2018). CHAI: Coding heuristics for assessing intuitive interaction. In *International Conference of Design, User Experience, and Usability* (S. 528–545). Heidelberg, Deutschland: Springer.

Repovš, G., & Baddeley, A. (2006). The multi-component model of working memory: Explorations in experimental cognitive psychology. *Neuroscience*, *139*(1), 5–21.

Resig, J. (2017). jQuery (Version 2.0.3) [Software]. Abgerufen von https://jquery.com/.

Ressing, M., Blettner, M., & Klug, S. J. (2009). Systematische Übersichtsarbeiten und Metaanalysen. *Deutsches Ärzteblatt International*, *106*(27), 456–63.

Rey, G. D. (2012). A review of research and a meta-analysis of the seductive detail effect. *Educational Research Review*, *7*(3), 216–237.

Rey, G. D. (2014). Seductive details and attention distraction – An eye tracker experiment. *Computers in Human Behavior*, *32*, 133–144.

Reyna, V. F. (2008). A theory of medical decision making and health: Fuzzy trace theory. *Medical Decision Making*, *28*(6), 850–865.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*(1), 1–75.

Ring, C., Burns, V. E., & Carroll, D. (2002). Shifting hemodynamics of blood pressure control during prolonged mental stress. *Psychophysiology*, *39*(5), 585–590.

Rizzo, L., Dondio, P., Delany, S. J., & Longo, L. (2016). Modeling mental workload via rule-based expert system: A comparison with NASA-TLX and workload profile. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* (S. 215–229). Springer. Heidelberg, Deutschland.

Robert McNeel & Associates. (2017). Rhinoceros (Version 5) [Software]. Abgerufen von https://www.rhino3d.com/.

Rockwell, R. C. (1975). Assessment of multicollinearity: The Haitovsky test of the determinant. *Sociological Methods & Research*, *3*(3), 308–320.

Rosch, J. L., & Vogel-Walcutt, J. J. (2013). A review of eye-tracking applications as tools for training. *Cognition, Technology & Work*, *15*(3), 313–327.

Roscoe, A. H. (1987). *In-flight assessment of workload using pilot ratings and heart rate* (AGARDgraph Nr. 282). Britania Airways Ltd. Luton, Vereinigtes Königreich.

Roscoe, A. H. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration? *Biological Psychology*, *34*(2-3), 259–287.

Rötting, M. (2001). *Parametersystematik der Augen-und Blickbewegungen für arbeitswissenschaftliche Untersuchungen.* Düren, Deutschland: Shaker.

Royer, J. M. (1979). Theories of the transfer of learning. *Educational Psychologist*, *14*(1), 53–69.

Rubin, D. C., & Umanath, S. (2015). Event memory: A theory of memory for laboratory, autobiographical, and fictional events. *Psychological Review*, *122*(1), 1.

Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive control of cognitive processes in task switching. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(4), 763.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology*, *53*(1), 61–86.

Rudy, J. W. (2009). Context representations, context functions, and the parahippocampal–hippocampal system. *Learning & Memory*, *16*(10), 573–585.

Rugg, M. D., & Coles, M. G. (1995). *Electrophysiology of mind: Event-related brain potentials and cognition.* Oxford, Vereinigtes Königreich: Oxford University Press.

Rumelhart, D. E. (2017). Schemata: The building blocks of cognition. In *Theoretical Issues in Reading Comprehension* (S. 33–58). Abingdon, Vereinigtes Königreich: Routledge.

Ryu, K., & Myung, R. (2005). Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics*, *35*(11), 991–1009.

Sachin, M. (2017). Captura (Version 6.0.1) [Software]. Abgerufen von https://mathewsachin.github.io/Captura/.

Saifoulline, P., & Hemberger, C. (2011). Kognitive Kernkompetenzen zum Aufbau fundierter mentaler Modelle für die Bearbeitung komplexer Planungsprobleme. *Journal Psychologie des Alltagshandelns/Psychology of Everyday Activity*, *4*(2), 31–44.

Salkind, N. J. (2010). *Encyclopedia of research design.* Thousand Oaks, CA, Vereinigte Staaten: Sage Publications.

Sanders, J. R. (2013). *Handbuch der Evaluationsstandards: die Standards des Joint Committee on Standards for Educational Evaluation.* Heidelberg, Deutschland: Springer.

Sarnikar, S., & Murphy, M. (2012). A Usability Analysis Framework for Healthcare Information Technology. *International Journal of Technology Diffusion (IJTD)*, *3*(4), 20–28.

Sarodnick, F., & Brau, H. (2006). *Methoden der Usability Evaluation.* Bern, Schweiz: Huber.

Saroyan, A. (1992). Differences in expert practice: A case from formative evaluation. *Instructional Science*, *21*(6), 451–472.

Sarris, V. (1990). *Methodologische Grundlagen der Experimentalpsychologie: Bd. 1; Erkenntnisgewinnung und Methodik der experimentellen Psychologie.* München, Deutschland: Ernst Reinhardt Verlag.

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the construct of usability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (S. 1609–1618). ACM. New York, NY, Vereinigte Staaten.

Savage, R. E., Wierwille, W. W., & Cordes, R. E. (1978). Evaluating the sensitivity of various measures of operator workload using random digits as a secondary task. *Human Factors*, *20*(6), 649–654.

Scavone, G. P. (2003). The Pipe: Explorations with breath control. In *Proceedings of the 2003 Conference on New Interfaces for Musical Expression* (S. 15–18). Singapur, Singapur: National University of Singapore.

Schellekens, J. M., Sijtsma, G. J., Vegter, E., & Meijman, T. F. (2000). Immediate and delayed after-effects of long lasting mentally demanding work. *Biological Psychology*, *53*(1), 37–56.

Schermelleh-Engel, K., & Schweizer, K. (2008). Multitrait-Multimethod-Analysen. In *Testtheorie und Fragebogenkonstruktion* (S. 325–341). Heidelberg, Deutschland: Springer.

Schlick, C. M., Winkelholz, C., Motz, F., Duckwitz, S., & Grandt, M. (2010). Complexity assessment of human–computer interaction. *Theoretical Issues in Ergonomics Science*, *11*(3), 151–173.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *6*(4), 147.

Schmidt-Kretschmer, M., & Blessing, L. (2006). Strategic aspects of design methodologies: Understood or underrated? In *DS 36: Proceedings DESIGN 2006, the 9th International Design Conference, Dubrovnik, Croatia.* DESIGN, Zagreb, Kroatien.

Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). A closer look at split visual attention in system-and self-paced instruction in multimedia learning. *Learning and Instruction*, *20*(2), 100–110.

Schneider, M. (2019). *Blickbasierte Beanspruchungsmessung: Entwicklung und Evaluation eines Kalibrierungssystems zur individuellen Bewertung der mentalen Beanspruchung in der Mensch-Technik-Interaktion.* Karlsruhe, Deutschland: KIT Scientific Publishing.

Scholkmann, F., Boss, J., & Wolf, M. (2012). An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals. *Algorithms*, *5*(4), 588–603.

Scholz, D. (2006). *Intuitivität von Mensch-Maschine-Systemen aus Benutzersicht* (Unveröffentlichte Diplomarbeit), Technische Universität Berlin, Berlin, Deutschland.

Schultheis, H., & Jameson, A. (2004). Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioral methods. In *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (S. 225–234). Springer. Heidelberg, Deutschland.

Schwalm, M. (2009). *Pupillometrie als Methode zur Erfassung mentaler Beanspruchungen im automotiven Kontext* (Unveröffentlichte Dissertation), Universität des Saarlandes, Saarbrücken, Deutschland.

Schwarz, N. (2002). *Feelings as information: Moods influence judgments and processing strategies.* Cambridge, Vereinigtes Königreich: Cambridge University Press.

Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology*, *14*(4), 332–348.

Schwarz, N. (2015). Metacognition. In *APA Handbook of Personality and Social Psychology, Volume 1: Attitudes and Social Cognition.* (S. 203–229). APA Handbooks in Psychology. Washington DC, Vereinigte Staaten: American Psychological Association.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *Journal of Personality and Social Psychology*, *45*(3), 513.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*(4), 570–582.

Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in Experimental Social Psychology*, *39*, 127–161.

Scott, B., & Neil, T. (2009). *Designing web interfaces: Principles and patterns for rich interactions.* Newton, MA, Vereinigte Staaten: O'Reilly.

Scriven, M. (1967). *The Methodology of Evaluation.* Washington DC, Vereinigte Staaten: American Educational Research Association.

Scriven, M. (1991). *Evaluation thesaurus.* Thousand Oaks, CA, Vereinigte Staaten: Sage Publications.

Scriven, M. (1996). Types of evaluation and types of evaluator. *Evaluation Practice*, *17*(2), 151–161.

Sears, A. (1997). Heuristic walkthroughs: Finding the problems without the noise. *International Journal of Human-Computer Interaction*, *9*(3), 213–234.

Sedlmeier, P., & Renkewitz, F. (2008). *Forschungsmethoden und Statistik in der Psychologie.* Hallbergmoos, Deutschland: Pearson Studium.

Seel, N. M. (1991). *Weltwissen und mentale Modelle.* Göttingen, Deutschland: Hogrefe Verlag.

Seifert, K. (2002). *Evaluation multimodaler Computer-Systeme in frühen Entwicklungsphasen* (Unveröffentlichte Dissertation), Technische Universität Berlin, Berlin, Deutschland.

Semmer, N., & Pfäfflin, M. (1978). *Interaktionstraining: Ein handlungstheoretischer Ansatz zum Training sozialer Fertigkeiten.* Weinheim, Deutschland: Beltz.

Seo, S. (2006). *A review and comparison of methods for detecting outliers in univariate data sets* (Unveröffentlichte Dissertation), University of Pittsburgh, Pittsburgh, PA, Vereinigte Staaten.

Serif Inc. (2017). Affinity Designer (Version 1.6.0.89) [Software]. Abgerufen von https://affinity.serif.com/.

Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007). Galvanic skin response (GSR) as an index of cognitive load. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems* (S. 2651–2656). ACM. New York, NY, Vereinigte Staaten.

Shiffler, R. E. (1988). Maximum Z scores and outliers. *The American Statistician*, *42*(1), 79–80.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127.

Shor, R. E. (1975). An auditory analog of the Stroop test. *Journal of General Psychology*, *93*, 281.

Simeral, E. J., & Branaghan, R. (1997). Comparative analysis of heuristic and usability evaluation methods. In *Proceedings of the 1997 44th Annual Conference of the Society for Technical Communication* (S. 307–309). Society for Technical Communication. Fairfax, VI, Vereinigte Staaten.

Simmons, J. P., & Nelson, L. D. (2006). Intuitive confidence: Choosing between intuitive and nonintuitive alternatives. *Journal of Experimental Psychology*, *135*(3), 409.

Simonton, D. K. (1980). Intuition and analysis: A predictive and explanatory model. *Genetic Psychology Monographs*, *102*, 3–60.

Sinclair, M., & Ashkanasy, N. M. (2005). Intuition: Myth or a decision-making tool? *Management Learning*, *36*(3), 353–370.

Sinclair, M., Ashkanasy, N. M., Chattopadhyay, P., & Boyle, M. V. (2002). Determinants of intuitive decision making in management: The moderating role of affect. In *Managing Emotions in the Workplace* (Bd. 9, S. 143–163). M. E. Sharpe. Armonk, NY, Vereinigte Staaten.

Singer, M., & Tiede, H. L. (2008). Feeling of knowing and duration of unsuccessful memory search. *Memory & Cognition*, *36*(3), 588–597.

Sirevaag, E. J., Kramer, A. F., Wickens, C. D., Reisweber, M., Strayer, D. L., & Grenell, J. F. (1993). Assessment of pilot performance and mental workload in rotary wing aircraft. *Ergonomics*, *36*(9), 1121–1140.

Skidmore, S. T., & Thompson, B. (2013). Bias and precision of some classical ANOVA effect sizes when assumptions are violated. *Behavior Research Methods*, *45*(2), 536–546.

Slepian, M. L., & Ambady, N. (2012). Fluid movement and creativity. *Journal of Experimental Psychology*, *141*(4), 625.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3.

Smith, D. M., & Mizumori, S. J. (2006). Hippocampal place cells, context, and episodic memory. *Hippocampus*, *16*(9), 716–729.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*(2), 108–131.

Sowa, J. F. (2000). *Knowledge representation: Logical, philosophical, and computational foundations*. Pacific Grove, CA, Vereinigte Staaten: Brooks/Cole.

Spchuang. (2016). VideoJS-Markers (Version 0.7.0) [Software]. Abgerufen von https://github.com/spchuang/videojs-markers.

Splawn, J. M., & Miller, M. E. (2013). Prediction of perceived workload from task performance and heart rate measures. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Bd. 57, *1*, S. 778–782). Sage Publications. Los Angeles, CA, Vereinigte Staaten.

Spool, J., & Schroeder, W. (2001). Testing web sites: Five users is nowhere near enough. In *CHI'01 Extended Abstracts on Human Factors in Computing Systems* (S. 285–286). New York, NY, Vereinigte Staaten: ACM.

Stanovich, K. E. (1999). *Who is rational?: Studies of individual differences in reasoning*. London, Vereinigtes Königreich: Psychology Press.

Stanovich, K. E. (2005). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago, IL, Vereinigte Staaten: University of Chicago Press.

Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory. In *Two Minds: Dual Processes and Beyond* (S. 55–88). Oxford, Vereinigtes Königreich: Oxford University Press.

Stanovich, K. E. (2011). *Rationality and the reflective mind*. Oxford, Vereinigtes Königreich: Oxford University Press.

Stanovich, K. E., & Toplak, M. E. (2012). Defining features versus incidental correlates of Type 1 and Type 2 processing. *Mind & Society*, *11*(1), 3–13.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2014). Rationality, intelligence, and the defining features of Type 1 and Type 2 processing. *Dual-process Theories of the Social Mind*, 80–91.

Stanton, N. (2016). On the reliability and validity of, and training in, ergonomics methods: A challenge revisited. *Theoretical Issues in Ergonomics Science*, *17*(4), 345–353.

Stanton, N., & Young, M. S. (1998). Is utility in the mind of the beholder? A study of ergonomics methods. *Applied Ergonomics*, *29*(1), 41–54.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, *57*(4), 421–457.

Still, J. D., Still, M. L., & Grgic, J. (2015). Designing intuitive interactions: Exploring performance and reflection measures. *Interacting with Computers*, *27*(3), 271–286.

Still, M. L., & Still, J. D. (2018). Cognitively describing intuitive interactions. In *Intuitive Interaction* (S. 41–61). Boca Raton, FL, Vereinigte Staaten: CRC Press.

Stillman, P. E., Shen, X., & Ferguson, M. J. (2018). How Mouse-tracking Can Advance Social Cognitive Theory. *Trends in Cognitive Sciences*, *22*(6), 531–543.

Stockmann, R. (2000). Evaluation in Deutschland. In *Evaluationsforschung* (S. 11–40). Heidelberg, Deutschland: Springer.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*(3), 220–247.

Sträter, O. (2016). *Cognition and safety: An integrated approach to systems design and assessment.* Abingdon, Vereinigtes Königreich: Routledge.

Strauss, A. L., Corbin, J. M., Niewiarra, S., & Legewie, H. (1996). *Grounded theory: Grundlagen qualitativer Sozialforschung.* Weinheim, Deutschland: Beltz.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643.

Stuiver, A., Brookhuis, K. A., De Waard, D., & Mulder, B. (2014). Short-term cardiovascular measures for driver support: Increasing sensitivity for detecting changes in mental workload. *International Journal of Psychophysiology*, *92*(1), 35–41.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring cognitive load. In *Cognitive Load Theory* (S. 71–85). Heidelberg, Deutschland: Springer.

Szulewski, A., Gegenfurtner, A., Howes, D. W., Sivilotti, M. L., & Van Merriënboer, J. J. (2017). Measuring physician cognitive load: Validity evidence for a physiologic and a psychometric tool. *Advances in Health Sciences Education*, *22*(4), 951–968.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics.* Bosten, MA, Vereinigte Staaten: Pearson.

Tattersall, A. J., & Hockey, G. R. J. (1995). Level of operator control and changes in heart rate variability during simulated flight maintenance. *Human Factors*, *37*(4), 682–698.

TechSmith. (2004). Morae (Version 3.3.4) [Software]. Abgerufen von https://www.techsmith.com/.

Thaler, R. H., & Shefrin, H. M. (1981). An economic theory of self-control. *Journal of political Economy*, *89*(2), 392–406.

Thomas, L. C., & Wickens, C. D. (2001). Visual displays and cognitive tunneling: Frames of reference effects on spatial judgments and change detection. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Bd. 45, *4*, S. 336–340). Sage Publications. Los Angeles, CA, Vereinigte Staaten.

Thompson, V. (2009). Dual process theories: A metacognitive perspective. *Ariel*, *137*, 51–43.

Thompson, V., Evans, J. S. B., & Campbell, J. I. (2013). Matching bias on the selection task: It's fast and feels good. *Thinking & Reasoning*, *19*(3), 431–452.

Thompson, V., & Morsanyi, K. (2012). Analytic thinking: do you feel like it? *Mind & Society*, *11*(1), 93–105.

Thompson, V., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140.

Thompson, V., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, *128*(2), 237–251.

Thurstone, L. L. (1944). *A factorial study of perception.* Chicago, IL, Vereinigte Staaten: University of Chicago Press.

Tidwell, J. (2010). *Designing interfaces: Patterns for effective interaction design.* Newton, MA, Vereinigte Staaten: O'Reilly.

Toates, F. (2006). A model of the hierarchy of behaviour, cognition, and consciousness. *Consciousness and Cognition*, *15*(1), 75–118.

Topolinski, S. (2011). A process model of intuition. *European Review of Social Psychology*, *22*(1), 274–315.

Topolinski, S., Likowski, K. U., Weyers, P., & Strack, F. (2009). The face of fluency: Semantic coherence automatically elicits a specific pattern of facial muscle reactions. *Cognition and Emotion*, *23*(2), 260–271.

Topolinski, S., & Sparenberg, P. (2012). Turning the hands of time: Clockwise movements increase preference for novelty. *Social Psychological and Personality Science*, *3*(3), 308–314.

Topolinski, S., & Strack, F. (2009). The architecture of intuition: fluency and affect determine intuitive judgments of semantic and visual coherence and judgments of grammaticality in artificial grammar learning. *Journal of Experimental Psychology*, *138*(1), 39.

Topolinski, S., & Strack, F. (2010). False fame prevented: Avoiding fluency effects without judgmental correction. *Journal of Personality and Social Psychology*, *98*(5), 721.

Tractinsky, N. (2018). The usability construct: A dead end? *Human–Computer Interaction*, *33*(2), 131–177.

Tracy, J. P., & Albers, M. J. (2006). Measuring cognitive load to test the usability of web sites. In *Society for Technical Communication. Annual Conference.* (Bd. 53, S. 256–260). Society for Technical Communication. Fairfax, VI, Vereinigte Staaten.

Trafton, J. G., & Monk, C. A. (2007). Task interruptions. *Reviews of Human Factors and Ergonomics*, *3*(1), 111–126.

Trejo, L. J., Kramer, A. F., & Arnold, J. A. (1995). Event-related potentials as indices of display-monitoring performance. *Biological Psychology*, *40*(1-2), 33–71.

Tretter, S., Diefenbach, S., & Ullrich, D. (2018). Intuitive Interaction from an Experiential Perspective: The Intuitivity Illusion and Other Phenomena. In *Intuitive Interaction* (S. 151–169). Boca Raton, FL, Vereinigte Staaten: CRC Press.

Trimble Navigation Ltd. (2016). SketchUp (Version 17.2.2554) [Software]. Abgerufen von https://www.sketchup.com.

Truijens, C., Trumbo, D., & Wagenaar, W. (1976). Amphetamine and barbiturate effects on two tasks performed singly and in combination. *Acta Psychologica*, *40*(3), 233–244.

Tsai, Y.-F., Viirre, E., Strychacz, C., Chase, B., & Jung, T.-P. (2007). Task performance and eye activity: Predicting behavior relating to cognitive workload. *Aviation, Space, and Environmental Medicine*, *78*(5), B176–B185.

Tsang, P. S. (2006). Regarding time-sharing with convergent operations. *Acta Psychologica*, *121*(2), 137–175.

Tsang, P. S., & Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, *39*(3), 358–381.

Tscharn, R., Latoschik, M. E., Löffler, D., & Hurtienne, J. (2017). "Stop over there": Natural gesture and speech interaction for non-critical spontaneous intervention in autonomous driving. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (S. 91–100). New York, NY, Vereinigte Staaten: ACM.

Tulving, E. (1985). *Elements of episodic memory.* Oxford, Vereinigtes Königreich: Oxford University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Ulich, E. (1994). *Arbeitspsychologie*. Stuttgart, Deutschland: Poeschel.

Ullrich, D. (2013). Komponenten und Einflussfaktoren der intuitiven Interaktion: Ein integratives Modell. *i-com*, *12*(3), 44–53.

Ullrich, D. (2014). *Intuitive Interaktion: Eine Exploration von Komponenten, Einflussfaktoren und Gestaltungsansätzen aus der Perspektive des Nutzererlebens* (Unveröffentlichte Dissertation), Technische Universität Darmstadt, Darmstadt, Deutschland.

Ullrich, D., & Diefenbach, S. (2010a). From magical experience to effortlessness: An exploration of the components of intuitive interaction. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (S. 801–804). New York, NY, Vereinigte Staaten: ACM.

Ullrich, D., & Diefenbach, S. (2010b). INTUI. Exploring the facets of intuitive interaction. In J. Ziegler & A. Schmidt (Hrsg.), *Mensch & Computer 2010: Interaktive Kulturen* (S. 251–260). München, Deutschland: Oldenbourg Verlag.

Ullrich, D., & Diefenbach, S. (2011). Erlebnis intuitive Interaktion — ein phänomenlogischer Ansatz. *i-com*, *10*(3), 63–68.

Ullsperger, M., & Von Cramon, D. Y. (2001). Subprocesses of performance monitoring: A dissociation of error processing and response competition revealed by event-related fMRI and ERPs. *Neuroimage*, *14*(6), 1387–1401.

Van den Haak, M., & De Jong, M. D. (2003). Exploring two methods of usability testing: Concurrent versus retrospective think-aloud protocols. In *Professional Communication Conference, 2003. IPCC 2003. Proceedings. IEEE International* (S. 285–287). Piscataway, NJ, Vereinigte Staaten: IEEE.

Van Den Haak, M., De Jong, M., & Jan Schellens, P. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology*, *22*(5), 339–351.

Van den Haak, M., de Jong, M. D., & Schellens, P. J. (2004). Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison. *Interacting with Computers*, *16*(6), 1153–1170.

Van Duyne, D. K., Landay, J. A., & Hong, J. I. (2007). *The design of sites: Patterns for creating winning web sites*. Upper Saddle River, NJ, Vereinigte Staaten: Prentice Hall Professional.

Van Welie, M., & Van der Veer, G. C. (2003). Pattern languages in interaction design: Structure and organization. In *Proceedings of INTERACT '03: IFIP TC13 International Conference on Human-Computer Interaction* (Bd. 3, S. 1–5). International Federation for Information Processing. Amsterdam, Niederlande.

Veenman, M. V., Van Hout-Wolters, B. H., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, *1*(1), 3–14.

Veltman, J., & Gaillard, A. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology*, *42*(3), 323–342.

Veltman, J., & Gaillard, A. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, *41*(5), 656–669.

VideoLAN. (2017). VLC media player (Version 2.2.8) [Software]. Abgerufen von https://www.videolan.org/.

Vidulich, M. A., & Tsang, P. S. (1985). Assessing subjective workload assessment: A comparison of SWAT and the NASA-bipolar methods. In *Proceedings of the Human Factors Society Annual Meeting* (Bd. 29, *1*, S. 71–75). Sage Publications. Los Angeles, CA, Vereinigte Staaten.

Vidulich, M. A., & Tsang, P. S. (2015). The confluence of situation awareness and mental workload for adaptable human–machine systems. *Journal of Cognitive Engineering and Decision Making*, *9*(1), 95–97.

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, *34*(4), 457–468.

Vollrath, M. (2015). *Ingenieurpsychologie: Psychologische Grundlagen und Anwendungsgebiete*. Stuttgart, Deutschland: Kohlhammer Verlag.

Volpert, W. (1982). The model of the hierarchical-sequential organization of action. In W. Hacker, W. Volpert, & M. von Cranach (Hrsg.), *Cognitive and motivational aspects of action.* Amsterdam, Niederlande: North Holland.

Walton, D. (2013). *Fallacies arising from ambiguity.* Heidelberg, Deutschland: Springer.

Wandke, H. (2004). Usability-testing. In *Lehrbuch der Medienpsychologie* (S. 325–354). Göttingen, Deutschland: Hogrefe Verlag.

Wegerich, A., Löffler, D., & Maier, A. (2012). *Handbuch zur IBIS Toolbox.* Berlin, Deutschland: Technische Universität Berlin.

Welford, A. (1967). Single-channel operation in the brain. *Acta Psychologica*, *27*, 5–22.

Wennecker, G. (2001). *Die Fokusgruppe. Eine Methode für die Requirementsanalyse zur ergonomischen Systemgestaltung.* (Unveröffentlichte Diplomarbeit), Universität Osnabrück, Osnabrück, Deutschland.

Wesson, J., & Cowley, L. (2003). Designing with patterns: Possibilities and pitfalls. In *Proceedings of the 2nd Workshop on Software and Usability Cross-Pollination: The Role of Usability Patterns.* International Federation for Information Processing, Amsterdam, Niederlande.

Whittlesea, B. W. (1993). Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(6), 1235.

Whittlesea, B. W., & Leboe, J. P. (2003). Two fluency heuristics (and how to tell them apart). *Journal of Memory and Language*, *49*(1), 62–79.

Wickens, C. D. (1980). The structure of attentional resources. *Attention and Performance*, *8*, 239–257.

Wickens, C. D. (1991). Processing resources and attention. *Multiple-Task Performance*, *1991*, 3–34.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177.

Wickens, C. D. (2004). Multiple resource time sharing models. In *Handbook of Human Factors and Ergonomics Methods* (S. 427–434). Boca Raton, FL, Vereinigte Staaten: CRC Press.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*(3), 449–455.

Wickens, C. D. (2017). Mental workload: Assessment, prediction and consequences. In *International Symposium on Human Mental Workload: Models and Applications* (S. 18–29). Springer. Heidelberg, Deutschland.

Wickens, C. D., & Alexander, A. L. (2009). Attentional tunneling and task management in synthetic vision displays. *The International Journal of Aviation Psychology*, *19*(2), 182–199.

Wickens, C. D., & Colcombe, A. (2007). Dual-task performance consequences of imperfect alerting associated with a cockpit display of traffic information. *Human Factors*, *49*(5), 839–850.

Wickens, C. D., Dixon, S. R., & Seppelt, B. (2005). Auditory preemption versus multiple resources: Who wins in interruption management? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Bd. 49, *3*, S. 463–466). Sage Publications. Thousand Oaks, CA, Vereinigte Staaten.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2015). *Engineering psychology & human performance*. London, Vereinigtes Königreich: Psychology Press.

Wickens, C. D., & Liu, Y. (1988). Codes and modalities in multiple resources: A success and a qualification. *Human Factors*, *30*(5), 599–616.

Widmer, T., & De Rocchi, T. (2012). *Evaluation: Grundlagen, Ansätze und Anwendungen*. Zürich, Schweiz: Rüegger Verlag.

Widyanti, A. (2017). Conceptual design of real time and adaptive measure of mental workload using galvanic skin respond. In *IOP Conference Series: Materials Science and Engineering* (Bd. 277, *1*, S. 012014). IOP Publishing. Bristol, Vereinigtes Königreich.

Wiebelitz, R., & Schmitz, H. (1983). Flimmerndes Licht und die menschliche Augenpupille. *Zeitschrift für Arbeitswissenschaft*, *37*(3), 163–168.

Wielgus, M. S., & Harvey, P. D. (1988). Dichotic listening and recall in schizophrenia and mania. *Schizophrenia Bulletin*, *14*(4), 689–700.

Wierwille, W. W., & Connor, S. A. (1983). Evaluation of 20 workload measures using a psychomotor task in a moving-base aircraft simulator. *Human Factors*, *25*(1), 1–16.

Wierwille, W. W., & Eggemeier, F. T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors*, *35*(2), 263–281.

Wierwille, W. W., & Gutmann, J. C. (1978). Comparison of primary and secondary task measures as a function of simulated vehicle dynamics and driving conditions. *Human Factors*, *20*(2), 233–244.

Wierwille, W. W., Rahimi, M., & Casali, J. G. (1985). Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediational activity. *Human Factors*, *27*(5), 489–502.

Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing*. Cambridge, MA, Vereinigte Staaten: Academic Press.

Wilson, G. F. (1992). Applied use of cardiac and respiration measures: Practical considerations and precautions. *Biological Psychology*, *34*(2-3), 163–178.

Wilson, G. F. (2002). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology*, *12*(1), 3–18.

Wilson, G. F. (2005). Operator functional state assessment for adaptive automation implementation. In *Biomonitoring for Physiological and Cognitive Performance during Military Operations* (Bd. 5797, S. 100–104). Bellingham, WA, Vereinigte Staaten: International Society for Optics und Photonics.

Wilson, G. F., & Russell, C. A. (2003). Operator functional state classification using multiple psychophysiological features in an air traffic control task. *Human Factors*, *45*(3), 381–389.

Wilson, T. D. (2004). *Strangers to ourselves*. Cambridge, MA, Vereinigte Staaten: Harvard University Press.

Winkielman, P., & Cacioppo, J. T. (2001). Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation elicits positive affect. *Journal of Personality and Social Psychology*, *81*(6), 989.

Winkielman, P., Schwarz, N., Fazendeiro, T., Reber, R. et al. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. *The Psychology of Evaluation: Affective Processes in Cognition and Emotion*, *189*, 217.

Winkler, A., Baumann, K., Huber, S., Tscharn, R., & Hurtienne, J. (2016). Evaluation of an application based on conceptual metaphors for social interaction between vehicles. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (S. 1148–1159). New York, NY, Vereinigte Staaten: ACM.

Wobbrock, J. O., & Kientz, J. A. (2016). Research Contributions in Human-Computer Interaction. *Interactions*, *23*(3), 38–44.

Wongupparaj, P., Kumari, V., & Morris, R. G. (2015). The relation between a multicomponent working memory and intelligence: The roles of central executive and short-term storage functions. *Intelligence*, *53*, 166–180.

Woywode, M., Mädche, A., Wallach, D., & Plach, M. (2012). *Gebrauchstauglichkeit von Anwendungssoftware als Wettbewerbsfaktor für kleine und mittlere Unternehmen (KMU): Abschlussbericht*. Universität Mannheim. Mannheim, Deutschland.

Wu, X., & Li, Z. (2013). Secondary Task Method for Workload Measurement in Alarm Monitoring and Identification Tasks. In *Cross-Cultural Design. Methods, Practice, and Case Studies* (S. 346–354). Springer. Heidelberg, Deutschland.

Yen, P.-Y., Sousa, K. H., & Bakken, S. (2014). Examining construct and predictive validity of the Health-IT Usability Evaluation Scale: Confirmatory factor analysis and structural equation modeling results. *Journal of the American Medical Informatics Association*, *21*(e2), e241–e248.

Ying, L., Fu, S., Qian, X., & Sun, X. (2011). Effects of mental workload on long-latency auditory-evoked-potential, salivary cortisol, and immunoglobulin A. *Neuroscience Letters*, *491*(1), 31–34.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17.

Young, M. S., & Stanton, N. (2002). Attention and automation: New perspectives on mental underload and performance. *Theoretical Issues in Ergonomics Science*, *3*(2), 178–194.

Zacher, H. (2017). Action Regulation Theory. In O. Braddick (Hrsg.), *Oxford Research Encyclopedia of Psychology*. New York, NY, Vereinigte Staaten: Oxford University Press.

Zacher, H., & Frese, M. (2018). Action regulation theory: Foundations, current knowledge, and future directions. In *The Sage handbook of Industrial, Work and Organizational Psychology* (Bd. 2, S. 80–102). Thousand Oaks, CA, Vereinigte Staaten: Sage Publications.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*(2), 151.

Zajonc, R. B., & Markus, H. (1982). Affective and cognitive factors in preferences. *Journal of Consumer Research*, *9*(2), 123–131.

Zandbergen, R. (2015). *Predicting persistency of usability problems based on error classification: A longitudnial study on improving mobility for the elderly* (Unveröffentlichte Master Thesis), University of Twente, Enschede, Niederlande.

Zander, T., Öllinger, M., & Volz, K. G. (2016). Intuition and insight: Two processes that build on each other or fundamentally differ? *Frontiers in Psychology*, *7*, 1395.

Zapf, D., Brodbeck, F. C., Frese, M., Peters, H., & Prümper, J. (1992). Errors in working with office computers: A first validation of a taxonomy for observed errors in a field setting. *International Journal of Human-Computer Interaction*, *4*(4), 311–339.

Zapf, D., Brodbeck, F. C., & Prümper, J. (1989). Handlungsorientierte Fehlertaxonomie in der Mensch-Computer Interaktion. *Zeitschrift für Arbeits-und Organisationspsychologie*, *33*(4), 178–187.

Zeiner, K. M., Laib, M., Schippert, K., & Burmester, M. (2016). Identifying Experience Categories to Design for Positive Experiences with Technology at Work. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (S. 3013–3020). ACM.

Zeitlin, L. R. (1995). Estimates of driver mental workload: A long-term field trial of two subsidiary tasks. *Human Factors*, *37*(3), 611–621.

Zelazo, P. D., Moscovitch, M., & Thompson, E. (2007). *The Cambridge handbook of consciousness*. Cambridge, Vereinigtes Königreich: Cambridge University Press.

Zempel, J. (2003). *Strategien der Handlungsregulation* (Unveröffentlichte Dissertation), Universität Gießen, Gießen, Deutschland.

Zentner, M., & Strauss, H. (2017). Assessing musical ability quickly and objectively: Development and validation of the Short-PROMS and the Mini-PROMS. *Annals of the New York Academy of Sciences*, *1400*(1), 33–45.

Zhang, J., Yu, X., & Xie, D. (2010). Effects of mental tasks on the cardiorespiratory synchronization. *Respiratory Physiology & Neurobiology*, *170*(1), 91–95.

Zijlstra, F. R. H. (1993). *Efficiency in Work Behaviour: A Design Approach for Modern Tools*. Delft, Niederlande: Delft University Press.

Zijlstra, F. R. H., & Van Doorn, L. (1985). *The construction of a subjective effort scale*. Delft University of Technology. Delft, Niederlande.

Zimmerman, P. H., Bolhuis, J. E., Willemsen, A., Meyer, E. S., & Noldus, L. P. (2009). The Observer XT: A tool for the integration and synchronization of multimodal signals. *Behavior research methods*, *41*(3), 731–735.

Anhang A. Studienübergreifende Zusatzinformationen

Dieser Anhang enthält studienübergreifende Zusatzmaterialien (z.B. Einverständniserklärung), die bei allen durchgeführten Experimenten zum Einsatz kamen. Dabei wurden diese Materialien nicht bei allen Experimenten in vollständig identischer Form verwendet. Da es jedoch bei der konkreten Ausgestaltung der Materialien nur geringe Unterschiede gab (z.B. Namen der getesteten Software-Anwendungen), werden die Materialien an dieser Stelle nur exemplarisch vorgestellt. Ferner wurde sich entschieden, die kryptischen Platzhalter des E-Mail-Templates, welches vom Probandensystem genutzt wurde, durch sprechende Bezeichnungen zu ersetzen. In den Materialien enthaltene vertrauliche Informationen wurden außerdem geschwärzt.

A.1 Rekrutierung der Versuchspersonen über das Probandensystem des Instituts für Mensch-Computer-Medien der Universität Würzburg (Studien 1, 2, 4, 5, 6, 7)

A.1.1 Exemplarische Beschreibung des fünften Experiments im Probandensystem

Das Ziel des Experiments ist es, Probleme und positive Aspekte verschiedener Software-Anwendungen (z.B. Hotelbuchung) ausfindig zu machen.

A.1.2 Exemplarische für das fünfte Experiment verwendete E-Mail, um Studierende auf die Studie im Probandensystem aufmerksam zu machen

Hallo <Name des Studierenden>!

Hiermit möchten wir Sie zu einem neuen Experiment einladen.

Es stehen die folgenden Termine zur Auswahl: <Mögliche Versuchstermine>

Wenn Sie teilnehmen möchten, können Sie sich unter dem folgenden Link anmelden:

<Link zur Studienanmeldung>

Ihr Versuchsleiter

A.2 Rekrutierung der Versuchspersonen über den E-Mail-Verteiler der Universität Stuttgart der Fachschaft Luft- und Raumfahrttechnik (Studie 3)

Einladung zur Teilnahme an Studie

Liebe/r Testteilnehmer/in,

mein Name ist Miriam Steller und ich studiere Informationsdesign an der Hochschule der Medien in Stuttgart. Im Rahmen meiner Bachelorarbeit führe ich eine Studie zur Untersuchung von Interaktionen in CAD-Programmen durch.

Dazu suche ich Testpersonen, welche **folgende Bedingungen erfüllen**:

- Maschinenbau-Studium oder vergleichbarer Studiengang oder Berufstätigkeit in diesem Bereich
- regelmäßige Nutzung von CAD-Software (erfahrener Umgang)

Wichtige Information:
Mit der Studie werden **die Produkte getestet** und **nicht Sie**! Durch Ihre Teilnahme können bisherige, als auch zukünftige Produkte hinsichtlich der Nutzungsqualität optimiert werden. Sie leisten somit einen wertvollen Beitrag zur Weiterentwicklung der Bedienoberflächen.

Testzeitraum und Dauer:
Der Testzeitraum ist voraussichtlich vom **25.09. - 07.10.2017**. Ein individueller Termin erfolgt per Absprache.
Die Studie dauert **ca. 1-1,5h** und findet im UX-Labor der Hochschule der Medien in Vaihingen statt.

Adresse:
Hochschule der Medien
Fakultät Information und Kommunikation
Nobelstraße 8
70569 Stuttgart

Kontakt:
Bei Interesse können Sie mich erreichen unter:
Mobilnummer: [redacted]
Email-Adresse: [redacted]

Ich würde mich sehr über Ihre Teilnahme an der Studie freuen!

Mit freundlichen Grüßen
[redacted]

A.3 Exemplarische für das fünfte Experiment verwendete Teilnehmerinformation

Lehrstuhl für Psychologische Ergonomie

[redacted]

Ansprechpartner für eventuelle Rückfragen:

[redacted]

Lehrstuhl für Psychologische Ergonomie

Allgemeine Teilnehmerinformation über die Untersuchung

Lehrstuhl für Psychologische Ergonomie

Titel der Studie: *Nutzertest verschiedener Software-Anwendungen*

Herzlich willkommen bei unserer Studie zum "*Nutzertest verschiedener Software-Anwendungen*"! Wir danken Ihnen für Ihr Interesse an dieser Studie.

Wir untersuchen mit dieser Studie, welche Probleme und positiven Aspekte mit der Bedienung einer Software-Anwendung (Regalplanungssoftware oder Hotelbuchungswebsite) zusammenhängen.

Ablauf der Studie

Das folgende Experiment besteht aus einer Testphase und einem anschließenden retrospektiven Interview. Insgesamt dauert das Experiment 1 Stunde.

Ihre Aufgabe ist es während der Testphase eine Reihe von Aufgaben (z. B. Buchung eines Hotelzimmers) mit der Software **stillschweigend** und **ohne die Hilfe des Versuchsleiters** zu lösen. Der Versuchsleiter wird Ihnen zuvor die Aufgaben detailliert vorstellen. Bei der Lösung dieser Aufgaben wird Ihr **Bildschirminhalt** und **Ihre Hände auf Video** aufgezeichnet. Ihr Gesicht ist auf dem Video nicht zu sehen. Der Versuchsleiter wird Sie zusätzlich über seinen Monitor bei der Aufgabenbewältigung beobachten und sich Notizen machen. Unmittelbar nachdem Sie alle Aufgaben erledigt haben, wird sich der Versuchsleiter mit Ihnen zusammen das Videomaterial ansehen und Sie bitten zu erklären, was Sie auf dem Video machen. Auf diese Weise sollen Probleme und positive Aspekte der Software erkannt und deren Ursache dokumentiert werden. Das retrospektive Interview wird ebenfalls auf Video und zusätzlich auf Ton aufgezeichnet. Es dient als Backup, um durch das retrospektive Interview entstandene Unklarheiten beseitigen zu können. Des Weiteren werden von Ihnen während dem Experiment Alter, Geschlecht und Studiengang als demographische Daten erhoben.

Sollten Sie noch Fragen haben, wenden Sie sich damit bitte an den Versuchsleiter.

Freiwilligkeit und Anonymität

Die Teilnahme an der Studie ist freiwillig. Sie können jederzeit und ohne Angabe von Gründen die Teilnahme an dieser Studie beenden, ohne dass Ihnen daraus Nachteile entstehen. Auch wenn Sie die Studie vorzeitig abbrechen, haben Sie Anspruch auf *die entsprechenden Versuchspersonenstunden* für den bis dahin erbrachten Zeitaufwand.

Die im Rahmen dieser Studie erhobenen, oben beschriebenen Daten und persönlichen Mitteilungen werden vertraulich behandelt. So unterliegen diejenigen Projektmitarbeiter, die durch direkten Kontakt mit Ihnen über personenbezogene Daten verfügen, der Schweigepflicht. Des Weiteren wird die Veröffentlichung der Ergebnisse der Studie in anonymisierter Form erfolgen, d. h. ohne dass Ihre Daten Ihrer Person zugeordnet werden können.

Datenschutz

Die Erhebung und Verarbeitung Ihrer oben beschriebenen persönlichen Daten erfolgt pseudonymisiert im Lehrstuhl für Psychologische Ergonomie unter Verwendung einer Nummer und ohne Angabe Ihres Namens. Es existiert eine Kodierliste auf Papier, die Ihren Namen mit der Nummer verbindet. Die Kodierliste ist nur dem Versuchsleiter und dem Projektleiter zugänglich; das heißt, nur diese Personen können die erhobenen Daten mit Ihrem Namen in Verbindung bringen. Die Kodierliste wird in einem abschließbaren Schrank aufbewahrt und nach Abschluss der Datenauswertung, spätestens aber am 24.02.2017 vernichtet. Ihre Daten sind dann anonymisiert. Damit ist es niemandem mehr möglich, die erhobenen Daten mit Ihrem Namen in Verbindung zu bringen. Die anonymisierten Daten werden mindestens 10 Jahre gespeichert. Solange die Kodierliste existiert, können Sie die Löschung aller von Ihnen erhobenen Daten verlangen. Ist die Kodierliste aber erst einmal gelöscht, können wir Ihren Datensatz nicht mehr identifizieren. Deshalb können wir Ihrem Verlangen nach Löschung Ihrer Daten nur solange nachkommen, wie die Kodierliste existiert.

Vergütung

*Für die Teilnahme erhalten Sie **1 Versuchspersonenstunde** gutgeschrieben.*

A.4 Exemplarische für das fünfte Experiment verwendete Einverständniserklärung

Lehrstuhl für Psychologische Ergonomie

Lehrstuhl für Pschologische Ergonomie

Daniel Reinhardt

Ansprechpartner für eventuelle Rückfragen:

[redacted]

Einwilligungserklärung für Bild- und Tonaufnahmen

Lehrstuhl für Psychologische Ergonomie

Titel der Studie: *Nutzertest verschiedener Software-Anwendungen*

Ich (Name des Teilnehmers /der Teilnehmerin in Blockschrift)

_____

bin *schriftlich* von Herrn/Frau _____ darüber informiert worden, dass im Rahmen der Studie eine *Video- und Tonaufnahme* gemacht wird.

Die Aufnahme dient dazu, anhand des Videos den Bedienungsablauf bei der Nutzung der Software zusammen mit Ihnen zu besprechen und dadurch potentielle Probleme bzw. positive Aspekte in Zusammenhang mit der Software erkennen zu können. Des Weiteren kann bei Unstimmigkeiten bei der Auswertung mehrmals auf die Aufzeichnungen zurückgegriffen werden.

Ich bin darüber informiert, dass die Aufzeichnung und Auswertung der *Video- und Tonaufnahme pseudonymisiert erfolgt, d. h. unter Verwendung einer Nummer und ohne Angabe meines Namens und dass eine Kodierliste auf Papier existiert, die meinen Namen mit der Nummer verbindet. Die Kodierliste ist nur dem Versuchsleiter zugänglich und wird nach Abschluss der Datenanalyse gelöscht*. Es besteht die sehr geringe Wahrscheinlichkeit, dass eine an der Datenauswertung beteiligte Person mich erkennt. Aus diesem Grund unterliegen alle an der Auswertung beteiligten Personen einer absoluten Schweigepflicht und dürfen unter keinen Umständen vertrauliche Informationen an Dritte weitergeben.

Mir ist bekannt, dass ich mein Einverständnis zur Aufbewahrung bzw. Speicherung dieser Daten widerrufen kann, ohne dass mir daraus Nachteile entstehen. Die *Video- und Tonaufnahme* wird in einem verschlossenen Schrank aufbewahrt. Ich bin darüber informiert worden, dass ich jederzeit eine Löschung meiner Aufnahmen verlangen kann*, solange die Kodierliste existiert*. Die Aufnahmen werden aber in jedem Fall nach Abschluss der Analyse vernichtet.

Mit der beschriebenen Handhabung der erhobenen Aufnahmen bin ich einverstanden.

Zusatz für Demonstrationen *Ich gebe mein Einverständnis, dass meine Video- und Tonaufnahme zu Demonstrationszwecken in teilnehmerbegrenzten Veranstaltungen (z. B. Lehrveranstaltungen) abgespielt werden. Zutreffendes bitte ankreuzen: O JA O NEIN.*

Die Einverständniserklärung für die *Video- und Tonaufnahme* ist freiwillig. Ich kann diese Erklärung jederzeit widerrufen. Im Falle einer Ablehnung oder eines Rücktritts entstehen für mich keinerlei Kosten oder anderweitige Nachteile; eine Teilnahme an der Studie ist *dann allerdings nicht* möglich.

Vorlage der Ethikkommission der Deutschen Gesellschaft für Psychologie für die Einwilligungserklärung für Bild- und Tonaufnahmen
14. Januar 2016

Ich hatte genügend Zeit für eine Entscheidung. Ich habe alles gelesen und erkläre mich hiermit bereit, dass eine *Video- und Tonaufnahme* von mir gemacht wird.

Eine Ausfertigung dieser Einwilligungserklärung habe ich erhalten.

Zusatzvereinbarung für künftige Kontaktaufnahmen im Rahmen dieser Studie

Ich gebe mein Einverständnis, dass im Falle einer Fortführung dieser Studie oder von Anschlussstudien meine personenbezogenen Daten weiter verwendet werden dürfen. Dies dient einer erneuten Kontaktaufnahme zu mir im Rahmen dieser Studie. Ich bin darüber informiert, dass meine Daten bis zum endgültigen Abschluss der Datenerhebung und/oder Auswertung weiterhin in pseudonymisierter Form (Kodierliste) vorliegen und nur die Studienleitung darauf Zugriff hat. Nach spätestens 10 Jahren werden meine personenbezogenen Daten gelöscht. Bis dahin kann ich jederzeit Auskunft über meine personenbezogenen Daten erhalten und die Löschung dieser Daten verlangen.

O JA O NEIN.

Rückmeldung von Ergebnissen

Ich bin daran interessiert, etwas über die Ergebnisse der Studie zu erfahren, und bitte hierzu um Übersendung entsprechender Informationen.

O JA O NEIN.

Ort, Datum & Unterschrift des Teilnehmers: Name des Teilnehmers in Druckschrift:

Ort, Datum & Unterschrift des Versuchsleiters: Name des Versuchsleiters in Druckschrift:

Bei Fragen oder anderen Anliegen kann ich mich an folgende Personen wenden:

Versuchsleiter:	Projektleiter:
[redacted]	*Daniel Reinhardt, Lehrstuhl für Psychologische Ergonomie* *Oswald-Külpe-Weg 82, 97074 Würzburg* *0931 31 80355* *daniel.reinhardt@uni-wuerzburg.de*

Vorlage der Ethikkommission der Deutschen Gesellschaft für Psychologie für die Einwilligungserklärung für Bild- und Tonaufnahmen
14. Januar 2016

Anhang B. Studienspezifische Zusatzinformationen

Dieser Anhang enthält für jedes durchgeführte Experiment studienspezifische Zusatzmaterialien. Es wurde sich aufgrund größerer Unterschiede zwischen den Experimenten (z.B. getestete Software hat große Auswirkung auf die Items des eingesetzten TFQ) im Gegensatz zum vorherigen Anhang dagegen entschieden, lediglich exemplarische Beschreibungen der Materialien zu verwenden. Stattdessen wurden konkret, die in den jeweiligen Experimenten verwendeten, Materialien abgedruckt, da nur so eine mögliche Replikation gewährleistet werden kann. Demzufolge wurden auch die Formatierungen (z.B. fett gedruckter Text) und Formulierungen der Instruktionen ohne Änderungen übernommen. Da die verwendeten TFQs in den Experimenten lediglich beim ersten Experiment über LimeSurvey und ansonsten auf Papier administriert wurden, wurde sich auch beim ersten Experiment entschieden, den TFQ schriftlich und nicht in Form eines Screenshots darzustellen.

B.1 Experiment 1

B.1.1 Aufgaben

Aufgabe 1:
Suchen Sie nach einer Möglichkeit 3D-Objekte im Raum positionieren zu können. Positionieren Sie anschließend die Lampe auf der Kommode rechts neben der Tür, sodass diese der Länge nach mittig auf der Kommode steht.

Aufgabe 2:
Suchen Sie nach einer Möglichkeit 3D-Objekte im Raum rotieren zu können. Rotieren Sie anschließend den Stuhl um 180 Grad.

Aufgabe 3:
Suchen Sie nach einer Möglichkeit die Länge von 3D-Objekten im Raum messen zu können. Messen Sie anschließend die Höhe der Tür (gemessen von der oberen Türkante bis zur Bodenkante).

B.1.2 Technology Familiarity Questionnaire und soziodemographischer Fragebogen

VP-Nr.:

Teilnehmerauskunft

1. Alter:

_____

2. Geschlecht:

_____

3. Studiengang/Beruf:

_____

4. Wie häufig verwenden Sie die folgenden Software-Anwendungen (Wenn Sie noch **nie** eine derartige Anwendung **verwendet** haben, kreuzen Sie bitte **„Niemals“** an. Ansonsten kreuzen Sie bitte die Antwortalternative an, welche am ehesten Ihrer Einschätzung entspricht.)?

Anwendung	täglich	Mehrmals pro Woche	Ein bis zweimal pro Woche	Alle paar Wochen	Alle paar Monate	Lediglich einmal oder zweimal verwendet	Niemals
SketchUp							
AutoCAD							
Fusion 360							
SolidWorks							
Andere CAD-Software **Welche verwendest du?** _____							

5. Wenn Sie die folgenden Software-Anwendungen verwenden, wie viele Features nutzten Sie davon (Wenn Sie keine derartige Anwendung verwenden, kreuzen Sie bitte „Keine“ an. Ansonsten kreuzen Sie bitte die Antwortalternative an, welche am ehesten Ihrer Einschätzung entspricht.)?

Anwendung	Alle Features (du schaust dazu im Internet, Foren oder nutzt die Hilfefunktion)	So viele Features wie du ohne externe Hilfe entdecken kannst	Genug Features, um damit arbeiten zu können	Dein begrenztes Wissen über die verfügbaren Features schränkt deine Nutzung der Anwendung ein	Keines der Features – du verwendest die Anwendung nicht
SketchUp					
AutoCAD					
Fusion 360					
SolidWorks					
Andere CAD-Software **Welche verwendest du?** _____					

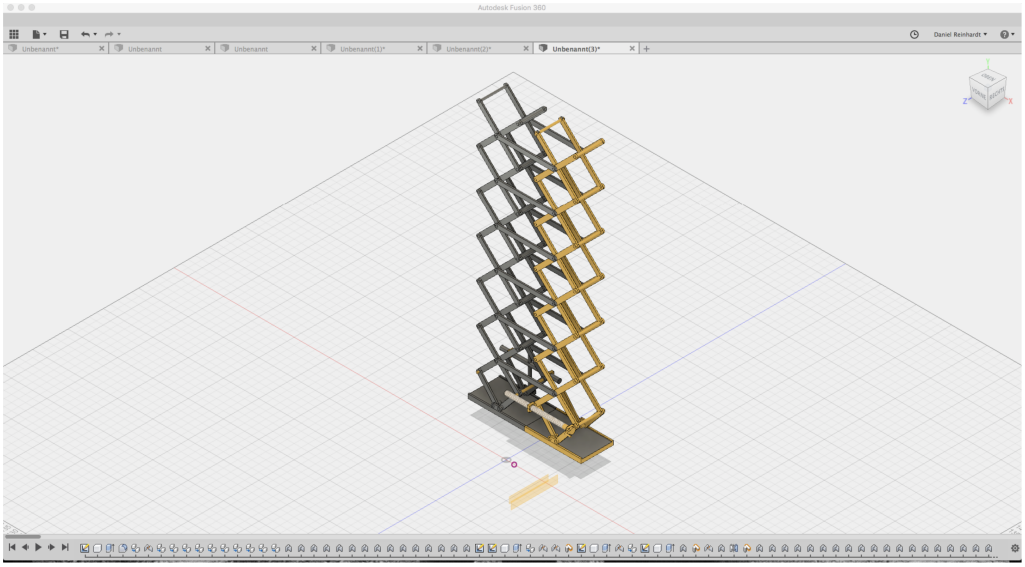
B.2 Experiment 2

B.2.1 Aufgaben

B.2.1.1 Fusion 360

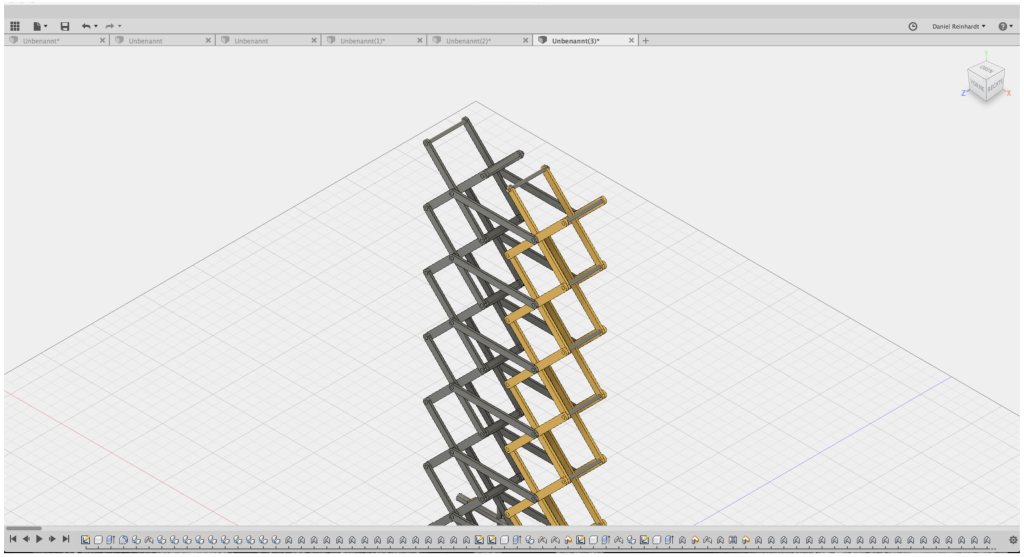
Nach dem Abschluss Ihres Maschinenbaustudiums finden Sie Ihre erste Anstellung als Ingenieur in der Anlagenplanung bei der Robert Bosch GmbH. Aus Gründen der Effizienz setzt Bosch CAD-Software zur virtuellen Planung ein. Ein Kollege hat Sie gebeten, die Farben eines von ihm erstellten Gerüsts anzupassen.

Aufgabe 1:



Zunächst soll die Farbe der Streben der rechten Gerüstseite angepasst werden. Verwenden Sie zur Selektion der Gerüststreben die Farbauswahl, welche Sie in der Menüleiste unter „Auswählen“ finden. Achten Sie bitte beim Auswählen der Streben darauf, dass nur die rechten Streben ausgewählt werden und nicht die mittleren Streben (siehe Bild). Nach der Auswahl klicken Sie auf „Rechtsklick“ und öffnen das Darstellungsmenü. Hier wählen Sie eine gelbe Farbe aus und färben die selektierten Streben ein.

Aufgabe 2:



Da nur Querstreben später eingefärbt sein sollen, müssen die anderen Streben (siehe Bild) wieder grau gemacht werden. Suchen Sie eine Möglichkeit dieses Streben zu markieren und dann wie bei Aufgabe 1 über das Darstellungsmenü einzufärben.

B.2.1.2 Affinity Designer

Dein Onkel hat nächste Woche Geburtstag. Um ihm eine Freude zu bereiten möchtest du ihm eine selbst gestaltete Geburtstagskarte zukommen lassen. Da du weißt, dass er ein riesiger Star Wars Fan ist möchtest du ihm eine Karte im Star Wars Look gestalten.

Aufgabe 1:
Um die Karte zu gestalten, lege ein neues Dokument an. Wähle für dieses die Größe A5 im Querformat. Füge anschließend das Bild von BB-8, den Lieblingscharakters deines Onkels, welches auf dem Desktop unter dem Namen „BB8.png“ gespeichert ist, zu deiner Karte hinzu. Öffne das Bild, sodass es in deinem Dokument erscheint und platziere es auf der linken Seite deiner Karte.

Aufgabe 2:
Füge drei Zahnräder mit unterschiedlicher Größe zu deiner Karte hinzu um diese zu dekorieren.

- Platziere das größte Zahnrad rechts unten und färbe es orange.
- Schließe die Öffnung in der Mitte des mittelgroßen Zahnrads und platziere es ungefähr mittig im Dokument.
- Setze nun das kleinste Zahnrad neben den Kopf von BB-8 und reduziere die Zahl der Zähne auf 8.
- Schreibe als Geburtstagsgruß „Happy BB-Day“ an den oberen Rand der Karte.
- Wähle die Schriftart „Tw Cen MT Condensed Extra Bold“.
- Färbe den Text gelb.
- Wähle die Größe 60pt.

Aufgabe 3:
Färbe zu guter Letzt den Hintergrund der Karte schwarz

B.2.2 Technology Familiarity Questionnaire und soziodemographischer Fragebogen

VP-Nr.:

Teilnehmerauskunft

1. Alter:

---

2. Geschlecht:

---

3. Studiengang/Beruf:

---

4. Wie häufig verwenden Sie die folgenden Software-Anwendungen (Wenn Sie noch **nie** eine derartige Anwendung **verwendet** haben, kreuzen Sie bitte **„Niemals“** an. Ansonsten kreuzen Sie bitte die Antwortalternative an, welche am ehesten Ihrer Einschätzung entspricht.)?

Anwendung	täglich	Mehrmals pro Woche	Ein bis zweimal pro Woche	Alle paar Wochen	Alle paar Monate	Lediglich einmal oder zweimal verwendet	Niemals
Affinity Designer							
Fusion 360							
Andere CAD-Software? **Welche verwendest du?** _____							

5. Wenn Sie die folgenden Software-Anwendungen verwenden, wie viele Features nutzten Sie davon (Wenn Sie keine derartige Anwendung verwenden, kreuzen Sie bitte „Keine“ an. Ansonsten kreuzen Sie bitte die Antwortalternative an, welche am ehesten Ihrer Einschätzung entspricht.)?

Anwendung	Alle Features (du schaust dazu im Internet, Foren oder nutzt die Hilfefunktion)	So viele Features wie du ohne externe Hilfe entdecken kannst	Genug Features, um damit arbeiten zu können	Dein begrenztes Wissen über die verfügbaren Features schränkt deine Nutzung der Anwendung ein	Keines der Features – du verwendest die Anwendung nicht
Affinity Designer					
Fusion 360					
Andere CAD-Sofware? **Welche verwendest du?** _____					

B.3 Experiment 3

B.3.1 Aufgaben

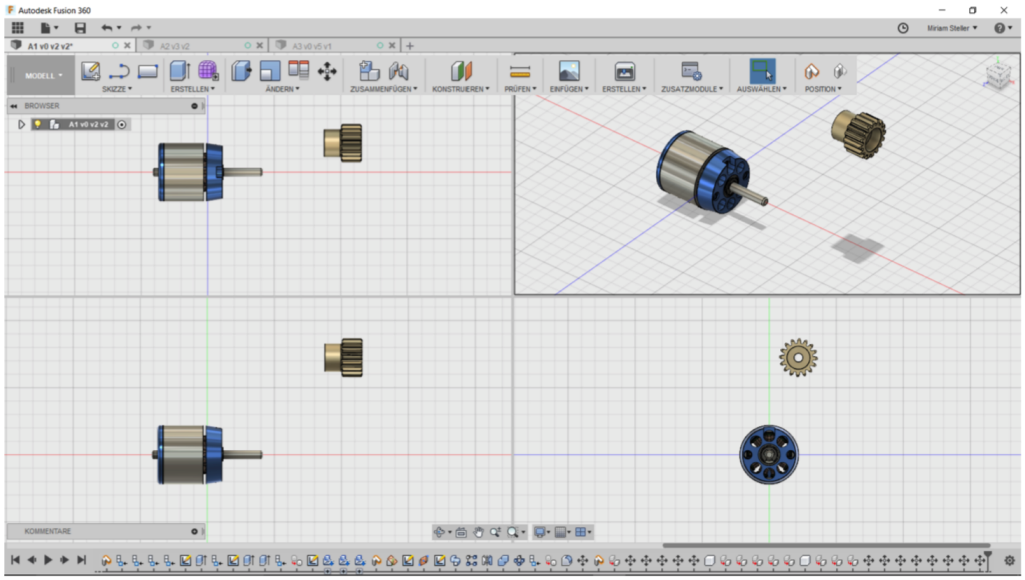
B.3.1.1 Fusion 360

Es wird die 3D-Konstruktionssoftware **Fusion 360** getestet. Mithilfe der Software kann man 2D- Zeichnungen sowie auch 3D-Objekte, also Bauteile aller Art, erstellen. Stelle Dir vor, Du arbeitest bei einem bekannten Modellauto-Hersteller in der Entwicklungsabteilung. Für das geplante neue Modellauto muss die Antriebseinheit, bestehend aus Motor und Ritzel, zusammengebaut werden. Vorgesehen ist dafür ein bürstenloser Elektromotor. Gearbeitet wird mit der 3D-Konstruktionssoftware **Fusion 360**. Da du in einer anstehenden wichtigen Präsentation eine vollständig zusammengesetzte Antriebseinheit zeigen möchtest, hast Du die notwendigen Schritte in 3 Teilaufgaben aufgeteilt, um den Überblick zu behalten.

Aufgabe 1: Rotieren

Drehe das Ritzel so, dass es wie auf dem Zielbild angezeigt im Raum liegt. Dabei soll das Ritzel nur um die eigenen Haupt-Achsen (X, Y, Z-Achse) gedreht werden. Verwende dazu das Tool *Verschieben/Kopieren*.

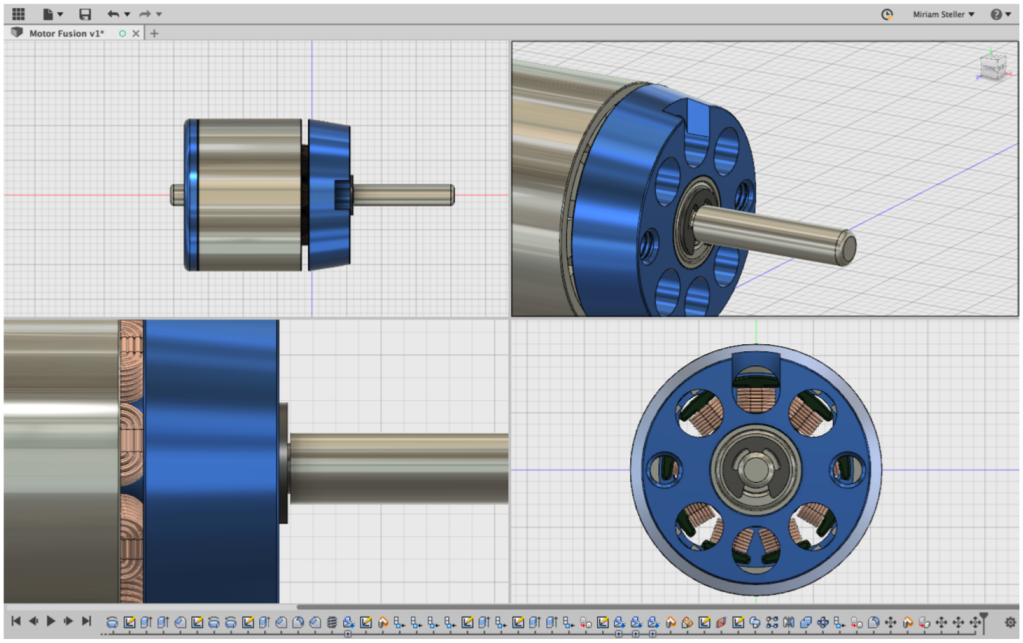
Das Zielbild zeigt, wie das Ritzel im Raum liegen soll.



Aufgabe 2: Snapping (Andocken)

Setze die Antriebswelle und den Stator so zusammen, dass die Nut am schwarzen Wellensicherungsring anliegt. Die beiden Körper sollen *keine* Einheit bilden. Verwende dazu das Tool *Ausrichten*.

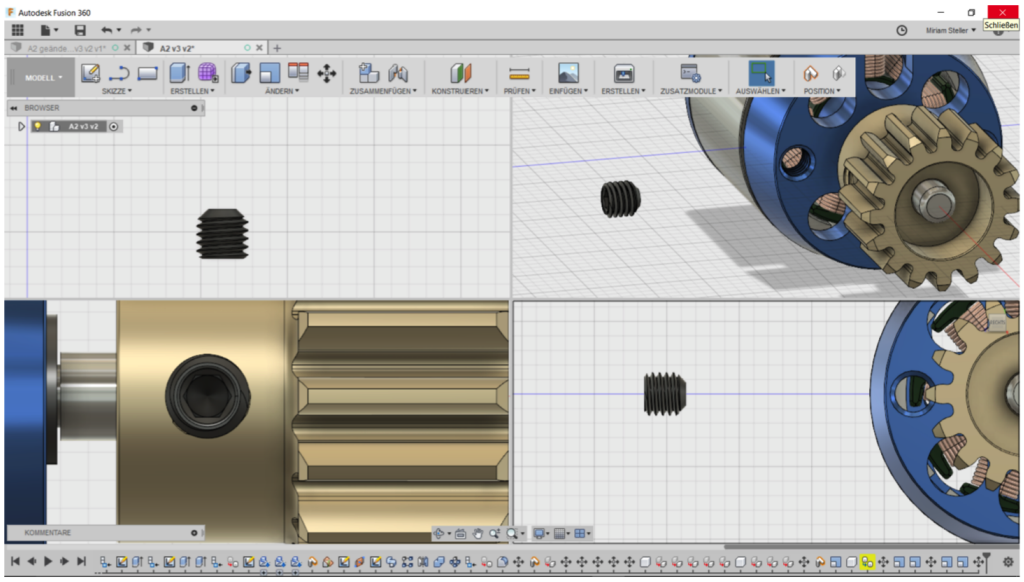
Das Zielbild zeigt die Antriebswelle mit dem Stator in zusammengesetzter Form.



Aufgabe 3: Skalieren

Du stellst fest, dass die Madenschraube nicht die richtigen Maße aufweist und findest heraus, dass die Schrauben-Länge halbiert und der Schraubenkopf auf die doppelte Größe verändert werden muss. Verwende dafür das Tool *Maßstab*. Dabei darfst du keine Zahlenwerte eingeben.

Das Zielbild zeigt, wie die Madenschraube nach Bearbeitung der Aufgabe aussehen soll.



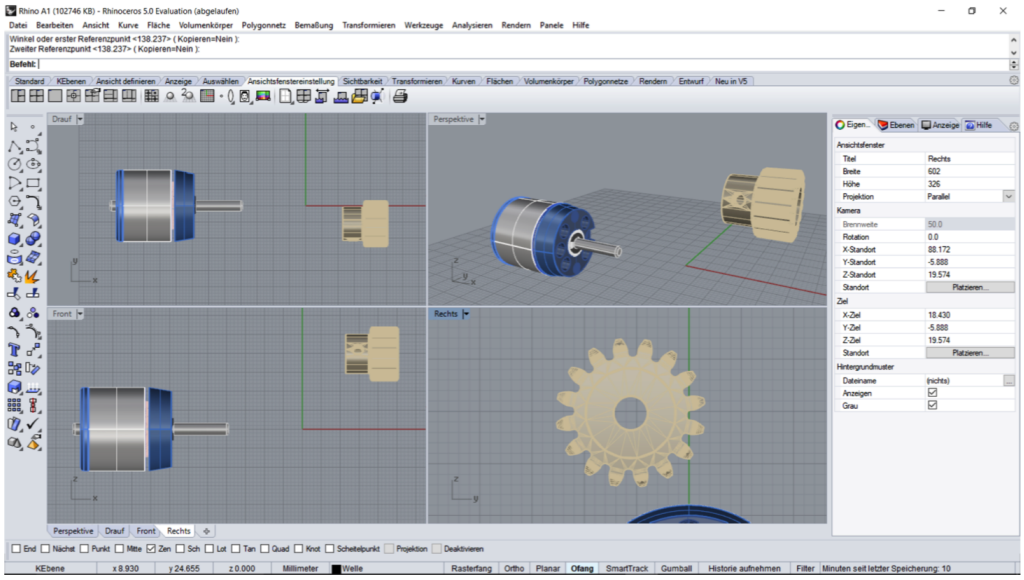
B.3.1.2 Rhinoceros 3D

Es wird die 3D-Konstruktionssoftware **Rhinoceros 3D** getestet. Mithilfe der Software kann man 2D- Zeichnungen sowie auch 3D-Objekte, also Bauteile aller Art, erstellen. Stelle Dir vor, Du arbeitest bei einem bekannten Modellauto-Hersteller in der Entwicklungsabteilung. Für das geplante neue Modellauto muss die Antriebseinheit, bestehend aus Motor und Ritzel, zusammengebaut werden. Vorgesehen ist dafür ein bürstenloser Elektromotor. Gearbeitet wird mit der 3D-Konstruktionssoftware **Rhinoceros 3D**. Da Du in einer anstehenden wichtigen Präsentation eine vollständig zusammengesetzte Antriebseinheit zeigen möchtest, hast Du die notwendigen Schritte in 3 Teilaufgaben aufgeteilt, um den Überblick zu behalten.

Aufgabe 1: Rotieren

Drehe das Ritzel so, dass es wie auf dem Zielbild angezeigt im Raum liegt. Dabei soll das Ritzel nur um die eigenen Haupt-Achsen (X, Y, Z-Achse) gedreht werden. Verwende dazu das Tool *Drehen oder Rotation*.

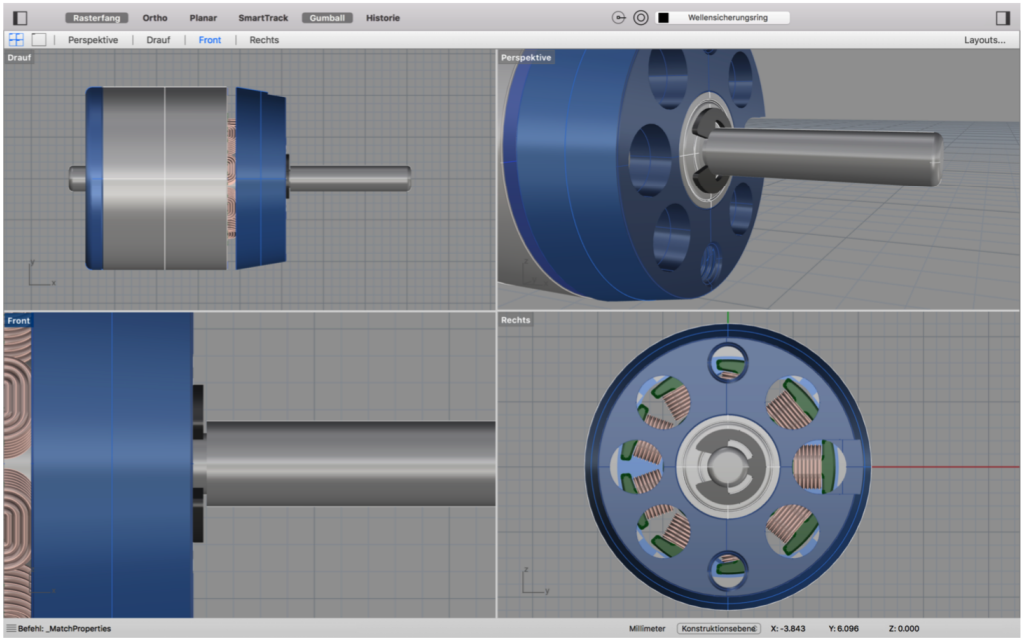
Das Zielbild zeigt, wie das Ritzel im Raum liegen soll.



Aufgabe 2: Snapping (Andocken)

Setze die Antriebswelle und den Stator so zusammen, dass die Nut am schwarzen Wellensicherungsring anliegt. Die beiden Bauteile sollen *keine* Einheit bilden. Verwende dazu das Tool *Verschieben*.

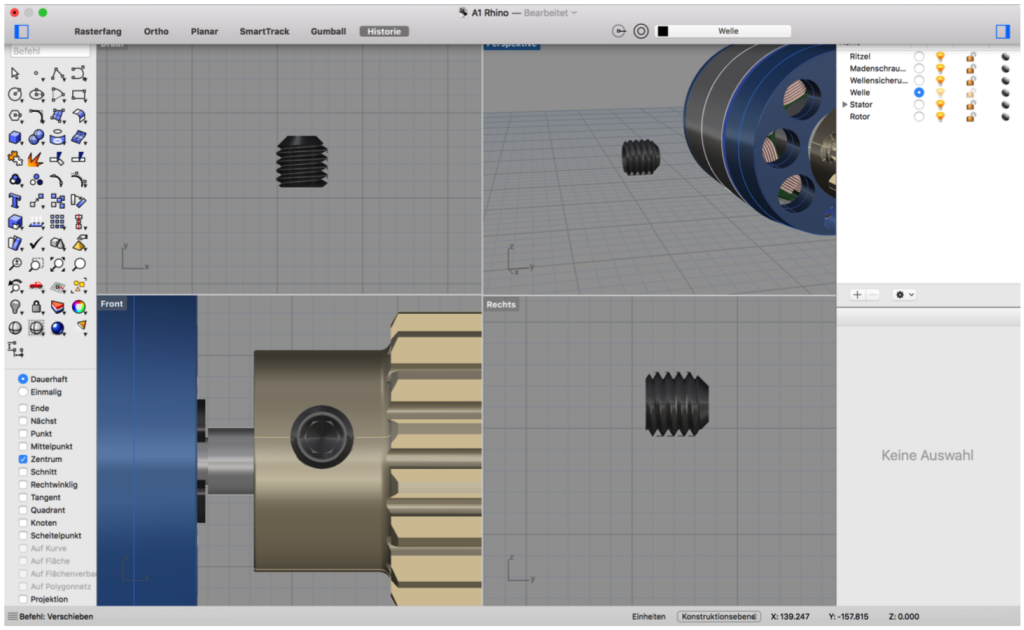
Das Zielbild zeigt die Antriebswelle mit dem Stator in zusammengesetzter Form.



Aufgabe 3: Skalieren

Du stellst fest, dass die Madenschraube nicht die richtigen Maße aufweist und findest heraus, dass die Schrauben-Länge halbiert und der Schraubenkopf-Durchmesser auf die doppelte Größe verändert werden muss. Verwende dafür das Tool *Skalieren*. Du darfst keine Zahlenwerte eingeben.

Das Zielbild zeigt, wie die Madenschraube nach Bearbeitung der Aufgabe aussehen soll.



B.3.2 Technology Familiarity Questionnaire und soziodemographischer Fragebogen

VP-Nr.:

Teilnehmerauskunft

1. Alter:

---

2. Geschlecht:

---

3. Studiengang/Beruf:

---

4. Wie häufig verwenden Sie die folgenden Software-Anwendungen (Wenn Sie noch **nie** eine derartige Anwendung **verwendet** haben, kreuzen Sie bitte **„Niemals“** an. Ansonsten kreuzen Sie bitte die Antwortalternative an, welche am ehesten Ihrer Einschätzung entspricht.)?

Anwendung	täglich	Mehrmals pro Woche	Ein bis zweimal pro Woche	Alle paar Wochen	Alle paar Monate	Lediglich einmal oder zweimal verwendet	Niemals
Fusion 360							
Rhinoceros 3D							
Andere CAD-Software **Welche verwendest du?** _____							

5. Wenn Sie die folgenden Software-Anwendungen verwenden, wie viele Features nutzten Sie davon (Wenn Sie keine derartige Anwendung verwenden, kreuzen Sie bitte „Keine“ an. Ansonsten kreuzen Sie bitte die Antwortalternative an, welche am ehesten Ihrer Einschätzung entspricht.)?

Anwendung	Alle Features (du schaust dazu im Internet, Foren oder nutzt die Hilfefunktion)	So viele Features wie du ohne externe Hilfe entdecken kannst	Genug Features, um damit arbeiten zu können	Dein begrenztes Wissen über die verfügbaren Features schränkt deine Nutzung der Anwendung ein	Keines der Features – du verwendest die Anwendung nicht
Fusion 360					
Rhinoceros 3D					
Andere CAD-Software **Welche verwendest du?** _____					

B.4 Experiment 4

B.4.1 Aufgaben

Sie arbeiten in der Fabrikplanung eines Großkonzerns in der Automobilbranche und sollen Regale für die Lagerung von Autoteilen mit der Software IPO.Rack planen.

Aufgabe 1: Festlegen von Regaltyp und Regalhersteller

- Bitte legen Sie den **Regaltyp 1** fest.
- Bitte wählen Sie das **Modell GRAPHIT** des **Herstellers TRILOGIQ** aus.
- Ihnen fällt ein, dass ihr Kunde doch lieber **Regaltyp 2** haben möchte und sie ändern es entsprechend.
- Da **TRILOGIQ** nicht das Modell GRAPHIT für diesen Typ anbietet, ändern sie das **Modell** auf **Lean Tek**.

Aufgabe 2: Festlegen verschiedener Regaldimensionen

- Ihr Kunde wünscht sich, dass das **Regal** eine **Höhe** von **2400 cm** aufweist, um Stoßstangen darin lagern zu können.
- Sie möchten das Regal mit **5 Regalböden** planen, die **jeweils** eine **Höhe** von **400 cm** haben.
- Alle von ihnen geplanten **Regalböden** sollen **1225 cm breit** sein.
- Sorgen Sie bitte dafür, dass die Stoßstangen später im Regal nicht von Regal rollen können indem Sie die **Anstellwinkel** der **einzelnen Regalböden** auf **0 %** setzen und alle Böden somit **waagrecht** sind.

Aufgabe 3: Anpassen des Aussehens des Regals

- Sie haben dem Kunden vom gerade vorliegende Regal einen Screenshot geschickt und ihm gefallen die Farben nicht. Ändern Sie deswegen diese folgendermaßen: Die **Fachböden** sollen **gelb** sein und der **Rahmen** soll **grün** sein.

Aufgabe 4: Stückliste (Einzelteile des Regals, die man benötigt um es zusammenzubauen) als PDF

- **Exportieren** Sie die **Stückliste** des vorliegenden Regals und legen sie die **PDF-Datei** auf Ihrem **Desktop** ab, um diese später Ihrem Lieferanten schicken zu können.

Aufgabe 5: Laden, Anpassen und Speichern eines bestehenden Regals

- **Laden** Sie bitte die **Regalkonfiguration** „AutoRegale“ (Sammlung mehrerer Regale).
- **Wähle** Sie bitte das Regal vom **Hersteller TRILOGIQ** aus, welches **4 Fächer besitzt** und **1400 hoch** ist. **Laden** Sie das Regal **anschließend**.
- Stoßstangen sollen später auf dem eben geladenen Regal platziert werden. Damit diese nicht verrutschen benötigen diese auf **jedem Regalboden zwei Rollbahnen** mit einem **Abstand** von **jeweils 500 cm**. Achten Sie beim Anlegen bitte darauf, dass Sie davor alle Regalböden komplett **waagrecht** stellen und alle **Anstellwinkel auf 0 %** stehen.
- **Legen** Sie für Ihr neues Regal eine neue Regalkonfiguration namens „Stossstangen-Regale“ an.
- Bitte **fügen** Sie Ihr **neues Regal** unter dem **Namen** „MiniStossstange“ Ihrer neuen **Konfiguration hinzu** und speichern das Ganze unter dieser **Regalkonfiguration** „StossstangenRegale“ ab.

B.4.2 Technology Familiarity Questionnaire und soziodemographischer Fragebogen

VP-Nr.:

Teilnehmerauskunft

1. Alter:

_____

2. Geschlecht:

_____

3. Studiengang/Beruf:

_____

4. Wie häufig verwenden Sie die folgenden Software-Anwendungen (Wenn Sie noch **nie** eine derartige Anwendung **verwendet** haben, kreuzen Sie bitte **„Niemals“** an. Ansonsten kreuzen Sie bitte die Antwortalternative an, welche am ehesten Ihrer Einschätzung entspricht.)?

Anwendung	täglich	Mehrmals pro Woche	Ein bis zweimal pro Woche	Alle paar Wochen	Alle paar Monate	Lediglich einmal oder zweimal verwendet	Niemals
IPO.Rack							
Andere Regalplanungssoftware **Welche verwendest du?** _____							

5. Wenn Sie die folgenden Software-Anwendungen verwenden, wie viele Features nutzten Sie davon (Wenn Sie keine derartige Anwendung verwenden, kreuzen Sie bitte „Keine“ an. Ansonsten kreuzen Sie bitte die Antwortalternative an, welche am ehesten Ihrer Einschätzung entspricht.)?

Anwendung	Alle Features (du schaust dazu im Internet, Foren oder nutzt die Hilfefunktion)	So viele Features wie du ohne externe Hilfe entdecken kannst	Genug Features, um damit arbeiten zu können	Dein begrenztes Wissen über die verfügbaren Features schränkt deine Nutzung der Anwendung ein	Keines der Features – du verwendest die Anwendung nicht
IPO.Rack					
Andere Regalplanungssoftware **Welche verwendest du?** _____					

B.5 Experiment 5

B.5.1 Aufgaben

Sie (Sasha Huber, sasha@byom.de) und Ihre gute Freundin (Vanessa Vogel, vanessa@byom.de) möchtet gerne **nächstes Jahr (20.03 - 04.04.2017)** zur **Kirschblüte** nach **Japan** reisen. Die **Flüge** habt ihr **schon gebucht** und jetzt müsst ihr euch nur noch für das passende Hotel entscheiden. Sie gehen dazu auf die Website Holidaycheck.de.

Aufgabe 1: Suchen eines passenden Zimmers in Tokyo. **Suchen** Sie für **sich** und ihre **gute Freundin** ein **Doppelzimmer** für den **20.03.2017 bis einschließlich 31.03.2017**. Das Zimmer soll die folgenden Merkmale aufweisen:

- WLAN muss vorhanden sein
- Hotelkategorie mind. 3 Sterne
- Frühstück sollte dabei sein
- Einchecken ab 06:00 Uhr morgens
- Kundenbewertung mind. 4 Sterne
- Der Preis für eine Nacht pro Person sollte 70 Euro nicht überschneiden
- Hotel und Zimmer sollten verfügbar sein

Ihre Aufgabe ist erfüllt, wenn ein passendes Zimmer gefunden wurde.

Aufgabe 2: Buchen eines passenden Zimmers in Tokyo

- Sie haben sich für das **Hotel Asia Center Japan in Tokyo** entschieden und möchten mit Ihrer guten Freundin, Vanessa Vogel, ein **Doppelzimmer buchen**. Wählen Sie dazu zunächst bitte das richtige Hotel und den Zeitraum vom **20.03.2017 bis einschließlich 31.03.2017** aus.
- Geben Sie im Anschluss bitte Ihre Personalien (Sasha Huber, 22.01.1986, Birkenweg 1, 97070 Würzburg, sasha@byom.de, 0981/7700) und die Personalien Ihrer guten Freundin (Vanessa Vogel, 13.08.1991, Semmelstraße 17, 97070 Würzburg, vanessa@byom.de, 0981/148621) ein.
- Sie möchten gerne das Ganze **selbst** per **Rechnung bezahlen** und sind an keine Reiseversicherung interessiert.
- Schließen Sie bitte Ihre Buchung ab. (**Achtung: Bestätigen Sie bitte nicht AGBs sonst wird das Hotel tatsächlich gebucht. Ihre Aufgabe endet sobald Sie die Buchung abschließen könnten.**)

Aufgabe 3: Hotel in Tokyo bewerten. Ihnen und Ihrer guten Freundin gefällt das **Hotel Asia Center Japan** besonders gut und möchten deswegen dem Hotel die **Bestnote** geben. Bewerte Sie dazu bitte das Hotel über die Website und berücksichtigen dabei die folgenden Punkte

- Preis-Leistung wurde als angemessen empfunden
- Sichere Weiterempfehlung kann ausgesprochen werden
- Zimmer haben Ihnen besonders gut gefallen
- Bei der Hotelbeschreibung (Freitext) schreiben Sie bitte folgendes: „Hotel hat eine super Lage und das Preis- und Leistungsverhältnis ist wirklich super. Das Frühstück ist sehr vielfältig und wir würden es wieder buchen."
- Die restlichen Angaben entnehmen Sie bitte dem Szenario.

Schließen Sie Ihre Bewertung jetzt ab. (**Achtung: Bewerten Sie bitte das Hotel nicht wirklich. Ihre Aufgabe endet sobald Sie die Bewertung abgeben könnten und alle benötigten Felder ausgefüllt sind!**)

Aufgabe 4: Hotel in Tokyo beobachten

- Um für Ihre bevorstehende Reise geeignete Hotels im Auge zu behalten und deren Preise zu überwachen, haben Sie sich bei Holidaycheck als Mitglied registriert und einen Login erhalten.
- Loggen Sie sich bitte zunächst in den geschützten Bereich von Holidaycheck ein.
- Anschließend füge Sie bitte das **Hotel Asia Center Japan Ihren Beobachtungen hinzu**, um es im Auge zu behalten und es nicht jedes Mal manuell suchen zu müssen. **Ihre Aufgabe ist danach abgeschlossen**.

Aufgabe 5: Nutzerprofil vervollständigen

- Um für Ihre bevorstehende Reise mit anderen Reisenden in Kontakt treten zu können, haben Sie sich bei Holidaycheck als Mitglied registriert und einen Login erhalten.
- **Logge** Sie sich bitte zunächst in den geschützten Bereich von Holidaycheck ein.
- Anschließend fügen Sie bitte Ihre **Traumreiseziele (Japan, Brasilien, Frankreich)** und Ihre bevorzugte Reiseart Ihrem Profil hinzu. Sie können so von Gleichgesinnten leichter gefunden werden.
- **Speichern** Sie bitte daraufhin Ihr **Profil** und beenden somit Ihre Aufgabe.

Aufgabe 6: Andere Reisende in Japan finden:

- Um für Ihre bevorstehende Reise mit anderen Reisenden in Kontakt treten zu können, haben Sie sich bei Holidaycheck als Mitglied registriert und einen Login erhalten.
- **Loggen** Sie sich bitte zunächst in den geschützten Bereich von Holidaycheck ein.
- Sie **möchten** sich im **Zeitraum** vom **01.04.2017 bis einschließlich 04.04.2017 gerne mit anderen Reisenden in Osaka treffen**. Dazu suchen Sie nach anderen Reisenden über Holidaycheck.
- Haben Sie einen anderen Reisenden gefunden, schreiben Sie ihm bitte über das **Portal** eine **Nachricht** mit dem folgenden Inhalt: „Hallo, haben Sie Lust mit mir zwischen dem 01.04 und 04.04 einen Kaffee in Tokyo zu trinken? LG, Sasha“.
- **Schicken** Sie bitte Ihre **Nachricht ab** und beenden somit Ihre Aufgabe.

B.5.2 Technology Familiarity Questionnaire und soziodemographischer Fragebogen

VP-Nr.:

Teilnehmerauskunft

1. Alter:

_____

2. Geschlecht:

_____

3. Studiengang/Beruf:

_____

4. Wie häufig verwenden Sie die folgenden Software-Anwendungen (Wenn Sie noch **nie** eine derartige Anwendung **verwendet** haben, kreuzen Sie bitte **„Niemals“** an. Ansonsten kreuzen Sie bitte die Antwortalternative an, welche am ehesten Ihrer Einschätzung entspricht.)?

Anwendung	täglich	Mehrmals pro Woche	Ein bis zweimal pro Woche	Alle paar Wochen	Alle paar Monate	Lediglich einmal oder zweimal verwendet	Niemals
Holidaycheck.de							
Andere Hotelbuchungswebsite **Welche verwendest du?** _____							

5. Wenn Sie die folgenden Software-Anwendungen verwenden, wie viele Features nutzten Sie davon (Wenn Sie **keine derartige Anwendung** verwenden, kreuzen Sie bitte **„Keine“** an. Ansonsten kreuzen Sie bitte die Antwortalternative an, welche am ehesten Ihrer Einschätzung entspricht.)?

Anwendung	Alle Features (du informierst dich dazu im Internet, Foren oder nutzt die Hilfefunktion)	So viele Features wie du ohne externe Hilfe entdecken kannst	Genug Features, um damit arbeiten zu können	Dein begrenztes Wissen über die verfügbaren Features schränkt deine Nutzung der Anwendung ein	Keines der Features – du verwendest die Anwendung nicht
Holidaycheck.de					
Andere Hotelbuchungswebsite **Welche verwendest du?** _____					

B.6 Experiment 6

B.6.1 Aufgaben

Sie haben sich einen 3D-Drucker gekauft und modellieren deswegen Ihr erstes 3D-Modell, um es später mit Ihrem neuen Drucker auszudrucken.

Aufgabe 1:
Fügen Sie der Arbeitsebene bitte ein Dach hinzu. Drehen Sie anschließend das Dach so, dass man es der Länge nach auf dem Quader platzieren könnte. Passen Sie nun die Größe des Dachs so an, dass es den Quader vollständig bedecken könnte.

Aufgabe 2:
Platzieren Sie das Dach auf dem Quader (justieren Sie ggf. Größe und Orientierung des Dachs nach).

B.6.2 Technology Familiarity Questionnaire und soziodemographischer Fragebogen

VP-Nr.:

Teilnehmerauskunft

1. Alter:

---

2. Geschlecht:

---

3. Studiengang/Beruf:

---

4. Wie häufig verwenden Sie die folgenden Software-Anwendungen (Wenn Sie noch **nie** eine derartige Anwendung **verwendet** haben, kreuzen Sie bitte **„Niemals“** an. Ansonsten kreuzen Sie bitte die Antwortalternative an, welche am ehesten Ihrer Einschätzung entspricht.)?

Anwendung	Täglich (6)	Mehrmals pro Woche	Ein bis zweimal pro Woche	Alle paar Wochen	Alle paar Monate	Lediglich einmal oder zweimal verwendet	Niemals
Tinkercad							
Andere CAD-Software? **Welche verwendest du?** _____							

5. Wenn Sie die folgenden Software-Anwendungen verwenden, wie viele Features nutzten Sie davon (Wenn Sie keine derartige Anwendung verwenden, kreuzen Sie bitte „Keine“ an. Ansonsten kreuzen Sie bitte die Antwortalternative an, welche am ehesten Ihrer Einschätzung entspricht.)?

Anwendung	Alle Features (du schaust dazu im Internet, Foren oder nutzt die Hilfefunktion)	So viele Features wie du ohne externe Hilfe entdecken kannst	Genug Features, um damit arbeiten zu können	Dein begrenztes Wissen über die verfügbaren Features schränkt deine Nutzung der Anwendung ein	Keines der Features – du verwendest die Anwendung nicht
Tinkercad					
Andere CAD-Sofware? **Welche verwendest du?** _____					

B.7 Experiment 7

B.7.1 Aufgaben

Aufgabe 1:
Du möchtest heute Mittag ein leckeres Risotto kochen. Hierfür fehlt dir noch Parmesan und Weißwein du beschließt, eine Einkaufsliste mit Hilfe von Chefkoch.de zu erstellen.

- Erstelle einen Einkaufszettel
- Benenne deine Liste in „Risotto“
- Füge Parmesan und Weißwein hinzu
- Benenne deine Liste in „Einkaufsliste_Risotto“ um

Aufgabe 2:
Deine Mitbewohnerin Anna und du habt euch bei „Running Dinner - Gemeinsam kochen“ angemeldet und sollt für heute Abend ein geeignetes Dessert kochen. Ein Blick in den Kühlschrank sagt euch, dass ihr noch Eier und ausreichend Butter übrig habt und beschließt, damit etwas zu backen. Da ein Teilnehmer keine Zartbitterschokolade mag, soll euer Rezept diese nicht enthalten. Um Rezepte zu finden, die zu eurem Vorrat passen, nutzt ihr das Portal Chefkoch.de.

Vorrätig hast du schon

- Butter
- Eier

Nicht rein darf

- Zartbitterschokolade

B.7.2 Technology Familiarity Questionnaire und soziodemographischer Fragebogen

VP-Nr.:

Teilnehmerauskunft

1. Alter:

_____

2. Geschlecht:

_____

3. Studiengang/Beruf:

_____

4. Wie häufig verwenden Sie die folgenden Software-Anwendungen (Wenn Sie noch **nie** eine derartige Anwendung **verwendet** haben, kreuzen Sie bitte **„Niemals"** an. Ansonsten kreuzen Sie bitte die Antwortalternative an, welche am ehesten Ihrer Einschätzung entspricht.)?

Anwendung	Täglich (6)	Mehrmals pro Woche	Ein bis zweimal pro Woche	Alle paar Wochen	Alle paar Monate	Lediglich einmal oder zweimal verwendet	Niemals
Chefkoch.de							
Anderes Webportal zum Thema kochen? **Welche verwendest du?** _____							

5. Wenn Sie die folgenden Software-Anwendungen verwenden, wie viele Features nutzten Sie davon (Wenn Sie keine derartige Anwendung verwenden, kreuzen Sie bitte „Keine" an. Ansonsten kreuzen Sie bitte die Antwortalternative an, welche am ehesten Ihrer Einschätzung entspricht.)?

Anwendung	Alle Features (du schaust dazu im Internet, Foren oder nutzt die Hilfefunktion)	So viele Features wie du ohne externe Hilfe entdecken kannst	Genug Features, um damit arbeiten zu können	Dein begrenztes Wissen über die verfügbaren Features schränkt deine Nutzung der Anwendung ein	Keines der Features – du verwendest die Anwendung nicht
Chefkoch.de					
Anderes Webportal zum Thema kochen? **Welche verwendest du?** _____					