Analyzing and fostering students' self-regulated learning through the use of peripheral data in online learning environments

Inaugural-Dissertation

zur Erlangung der Doktorwürde der

Fakultät für Humanwissenschaften

der

Julius-Maximilians-Universität Würzburg

Vorgelegt von

Markus Hörmann

aus Freiburg im Breisgau

Würzburg

2019

Erstgutachterin: Prof. Dr. Maria Bannert, Technische Universität München

Zweitgutachterin: Prof. Dr. Tina Seufert, Universität Ulm

Tag des Kolloquiums: 26.04.2019

Abstract

Learning with digital media has become a substantial part of formal and informal educational processes and is gaining more and more importance. Technological progress has brought overwhelming opportunities for learners, but challenges them at the same time. Learners have to regulate their learning process to a much greater extent than in traditional learning situations in which teachers support them through external regulation. This means that learners must plan their learning process themselves, apply appropriate learning strategies, monitor, control and evaluate it. These requirements are taken into account in various models of self-regulated learning (SRL). Although the roots of research on SRL go back to the 1980s, the measurement and adequate support of SRL in technology-enhanced learning environments is still not solved in a satisfactory way. An important obstacle are the data sources used to operationalize SRL processes. In order to support SRL in adaptive learning systems and to validate theoretical models, instruments are needed which meet the classical quality criteria and also fulfil additional requirements. Suitable data channels must be measurable "online", i.e., they must be available in real time during learning for analyses or the individual adaptation of interventions. Researchers no longer only have an interest in the final results of questionnaires or tasks, but also need to examine process data from interactions between learners and learning environments in order to advance the development of theories and interventions. In addition, data sources should not be obtrusive so that the learning process is not interrupted or disturbed. Measurements of physiological data, for example, require learners to wear measuring devices. Moreover, measurements should not be reactive. This means that other variables such as learning outcomes should not be influenced by the measurement. Different data sources that are already used to study and support SRL processes, such as protocols on thinking aloud, screen recording, eye tracking, log files, video observations or physiological sensors, meet these criteria to varying degrees. One data channel that has received little attention in research on educational psychology, but is non-obtrusive, non-reactive, objective and available online, is the detailed, timely high-resolution data on observable interactions of learners in online learning environments. This data channel is introduced in this thesis as "peripheral data". It records both the content of learning environments as context, and related actions of learners triggered by mouse and keyboard, as well as the reactions of learning environments, such as structural or content changes. Although the above criteria for the use of the data are met, it is unclear whether this data can be interpreted reliably and validly with regard to relevant variables and behavior.

Therefore, the aim of this dissertation is to examine this data channel from the perspective of SRL and thus further close the existing research gap. One development project and four research projects were carried out and documented in this thesis.

In the development work (chapter 4.1), "peripheral data" is described on a theoretical and methodological level and compared with the methods of screen recording, mouse/keyboard tracking and log files with regard to their advantages and disadvantages. Disadvantages of existing methods are for example the necessary manual coding of screen recording, the dependency on installed software for mouse/keyboard tracking, or the low granularity of log files. On a technical level, the development of a software framework called "ScreenAlytics" and its features for the acquisition and analysis of peripheral data is documented. In summary, researchers, not only from the field of educational psychology, can install the software on existing websites and record detailed data on interactions between users and web-based environments. ScreenAlytics uses this data to create video-like replays and other visualizations such as heat maps of mouse activity or navigation graphs. Replays can also be labelled with behavioral tags for qualitative analysis. Text input can be extracted with metadata for the writing process. An interface (API) allows researchers to export the collected

data in real time and record user-defined events. In addition, the system is evaluated and application scenarios are discussed.

The developed software framework forms the methodological basis for the following empirical studies, in which the relationships between this data source and SRL-relevant variables (learning success, motivation, cognitive load and confusion) are investigated. It also provides the data basis for the intervention study that examines the effects of learning dashboards and metacognitive prompts on learning outcomes.

The first empirical study (chapter 4.2) examined the relationship between typing behavior and learning outcomes, as well as current motivation. The rationale behind the recording and analysis of typing behavior is that the writing flow makes underlying cognitive processes observable. The analyses focus on various indices such as the length or frequency of pauses or corrections. The study assumes that indices of higher writing speed correlate positively with learning outcomes. With regard to motivation, Rheinberg and colleagues (2001) mention task processing time and the quality of task processing as potential indicators - this study assumes that both are also reflected in indices of the writing process. The study examined $N = 43$ students in an online learning environment for the acquisition of declarative and procedural knowledge about website programming. In the study, learners should first copy an example sentence to generate a baseline of typing behavior. Subsequently, initial motivation was collected with the Questionnaire of Current Motivation (QCM, Rheinberg, Vollmeyer & Burns, 2001), and spatial ability was measured with the VZ-2 Paper-Folding Test (Ekstrom, French, Harman & Dermen, 1976) as possible influencing variables. Moreover, declarative and procedural prior knowledge was measured. After the learners had worked on half of the learning environment, they were asked to write down their previous knowledge on three concepts in their own words in a recall task. They were also asked to solve two interactive tasks by writing programming code. After each of these tasks, a short version of the QCM

was presented in order to record the current motivation in high temporal proximity to the typing behavior. After learning, learners completed the same knowledge test that was taken before learning. Contrary to the hypothesis, it was found that indices of lower typing speed during the recall task correlate with both higher performance in the recall task and with higher values in the learning outcome test. Nevertheless, it could be shown that indices of higher typing speed during interactive programming tasks are associated with higher declarative and procedural knowledge acquisition. This pattern was also found with regard to current motivation. The findings of the context are discussed as representing task-specificity. During programming, the writing of a correct sequence of previously learned code chunks is required, which is why learners with fast and correct retrieval also show faster typing behavior. In contrast, writing continuous text requires the reconstruction and verbalization of knowledge. Slow writing or frequent corrections are interpreted rather as an expression of high standards. The discussion also addresses methodological problems and pending issues.

In the second study of this work (chapter 4.3), it was experimentally examined whether there is a connection between the mouse movements of learners and their CL and affective states. Mouse movement is regarded as a naturally occurring secondary task in the sense of the dual-task paradigm. The basic idea is that pauses in mouse movement occur when the load of the primary learning task is high. In a quasi-experimental study, $N = 49$ students were examined who learned online about website programming. Cognitive load was measured by reaction times. Learners had to press a key as quickly as possible when the background color of the learning environment changed. In addition, declarative and procedural knowledge as well as positive and negative affects were recorded with the PANAS instrument (Krohne, Egloff, Kohlmann & Tausch, 1996) prior to and after learning. In the experimental group ($N = 28$), the measurement of CL was only triggered when no mouse and keyboard input was registered for 6 seconds. In the control group ($N = 21$), the measurement was triggered at random intervals between 15 and 35 seconds. As assumed in the hypothesis, higher CL was observed in the experimental group with a medium effect ($d = .60$). In addition, significant correlations between mouse movements and affect could be shown for the control group. The results are very promising for a real-time measurement of these difficult to measure variables. The results, methodological limitations of the study and possible applications for interventions and further research are discussed.

The third study of this work (chapter 4.4) examined whether confusion, subjective and objective difficulty of items, as well as metacognitive assessments of one's own knowledge can be measured by mouse behavior. For this purpose, the mouse behavior of $N = 144$ persons was recorded when answering multi-item scales. Multi-Item scales were chosen because they follow a strict structure of question and answer options, but are still relevant for learning environments. Metacognitive Feeling-of-Knowing (FOK) judgements on 18 items of a crystalline intelligence test (BEFKI) were asked, followed by 60 items of a Big Five Inventory, of which 6 were manipulated with wrong grammar or contradictions to induce confusion. Afterwards, the actual answers to the BEFKI questions were acquired. It could be shown that 1) manipulated items can be recognized by increased indices of mouse behavior, 2) the strength of manipulation (contradiction > grammar) can be recognized by mouse behavior, 3) higher indices of mouse behavior are associated with higher subjective difficulty of items, but the power of this correlation is not sufficient to predict subjective difficulty, 4) higher indices of mouse behavior are associated with higher objective difficulty, but correlate low, 5) questions with higher FOK judgements have longer response times. In addition, detailed analyses were performed on various indices of mouse behavior. The study discusses the results taking into account existing evidence in the field of survey research. In addition, limitations of the study and possible applications of the results are discussed, including in rapid assessment tasks.

The fourth and final study of this work (chapter 4.5) examines the effects of learning dashboards to support the SRL and the detailed interaction of learners with this intervention. Learning dashboards contain visualizations of data about learning processes that have been previously collected, processed, and analyzed. The basic idea of the study is that learning dashboards support learning by providing information about the learning process and additional metacognitive prompts make this information relevant to learning strategies. The study implements recommendations from previous reviews. Thus, it was considered 1) to substantiate the intervention itself and the used data channels more theoretically on SRL frameworks, 2) to not only raise awareness about one's own learning process, but also to trigger changes of learning processes, as well as 3) to apply systematic experimental designs to investigate the effects of dashboards on learning outcome. The factors prompt and dashboard were experimentally varied. *N* = 138 learners were randomly distributed to a control group without intervention, a group with only prompts, a group with dashboards and a group with prompts and dashboards. Contents were the basics for programming JavaScript in an online learning environment. Learners were first shown short video trainings on how to use the interventions, then attitudes to privacy and metacognitive strategy knowledge were collected as covariates. A declarative and procedural knowledge test was followed by a 60-minute learning phase. After 20 and 40 minutes the respective intervention was presented. This was followed by the same knowledge test, an evaluation of the dashboard and a self-report on CL. There were no significant differences between the groups in terms of learning outcomes. The main reason given for this was the lack of need for regulation due to an excessively high predetermined structure of the learning environment. The detailed use of the interventions as well as the resulting CL and the perceived usefulness of different parts of the dashboard are discussed in detail. Recommendations for further research are also made.

Following the presentation of the five research and development projects, these will be considered in their overall context and the use of peripheral data in technology-based learning environments will be critically reflected.

Zusammenfassung (German Abstract)

Lernen mit digitalen Medien ist ein substantieller Bestandteil formeller und informeller Bildungsprozesse geworden und gewinnt noch immer an Bedeutung. Technologischer Fortschritt hat überwältigende Möglichkeiten für Lernende geschaffen, stellt aber gleichzeitig auch große Anforderungen an sie. Lernende müssen ihren Lernprozess sehr viel stärker selbst regulieren als in traditionellen Lernsituationen, in denen Lehrende durch externe Regulation unterstützen. Das heißt, Lernende müssen ihren Lernprozess selbst planen, geeignete Lernstrategien anwenden, ihn überwachen, steuern und evaluieren. Diesen Anforderungen wird in verschiedenen Modellen des selbst-regulierten Lernens (SRL) Rechnung getragen. Obwohl die Wurzeln der Forschung zu SRL bis in die 1980er Jahren zurück reichen, ist die Messung und adäquate Unterstützung von SRL in technologie-gestützten Lernumgebungen noch immer nicht zufriedenstellend gelöst. Eine wichtige Hürde sind dabei die Datenquellen, die zur Operationalisierung von SRL-Prozessen herangezogen werden. Um SRL in adaptiven Lernsystemen zu unterstützen und theoretische Modelle zu validieren, werden Instrumente benötigt, die klassischen Gütekriterien genügen und darüber hinaus weitere Anforderungen erfüllen. Geeignete Datenkanäle müssen „online“ messbar sein, das heißt bereits während des Lernens in Echtzeit für Analysen oder die individuelle Anpassung von Interventionen zur Verfügung stehen. Forschende interessieren sich nicht mehr nur für die Endergebnisse von Fragebögen oder Aufgaben, sondern müssen auch Prozessdaten von Interaktionen zwischen Lernenden und Lernumgebungen untersuchen, um die Entwicklung von Theorien und Interventionen voranzutreiben.

Zudem sollten Datenquellen nicht intrusiv sein, sodass der Lernprozess nicht unterbrochen oder gestört wird. Dies ist zum Beispiel bei Messungen physiologischer Daten der Fall, zu deren Erfassung die Lernenden Messgeräte tragen müssen. Außerdem sollten Messungen nicht reaktiv sein – andere Variablen (z.B. der Lernerfolg) sollten also nicht von der Messung

beeinflusst werden. Unterschiedliche Datenquellen die zur Untersuchung und Unterstützung von SRL-Prozessen bereits verwendet werden, wie z.B. Protokolle über lautes Denken, Screen-Recording, Eye Tracking, Log-Files, Videobeobachtungen oder physiologische Sensoren erfüllen diese Kriterien in jeweils unterschiedlichem Ausmaß. Ein Datenkanal, dem in der pädagogische-psychologischen Forschung bislang kaum Beachtung geschenkt wurde, der aber nicht-intrusiv, nicht-reaktiv, objektiv und online verfügbar ist, sind detaillierte, zeitlich hochauflösende Daten über die beobachtbare Interkation von Lernenden in online Lernumgebungen. Dieser Datenkanal wird in dieser Arbeit als „peripheral data" eingeführt. Er zeichnet sowohl den Inhalt von Lernumgebungen als Kontext auf, als auch darauf bezogene Aktionen von Lernenden, ausgelöst durch Maus und Tastatur, sowie die Reaktionen der Lernumgebungen, wie etwa strukturelle oder inhaltliche Veränderungen. Zwar sind die oben genannten Kriterien zur Nutzung der Daten erfüllt, allerdings ist unklar, ob diese Daten auch reliabel und valide hinsichtlich relevanten Variablen und Verhaltens interpretiert werden können.

Ziel dieser Dissertation ist es daher, diesen Datenkanal aus Perspektive des SRL zu untersuchen und damit die bestehende Forschungslücke weiter zu schließen. Dafür wurden eine Entwicklungs- sowie vier Forschungsarbeiten durchgeführt und in dieser Arbeit dokumentiert.

In der Entwicklungsarbeit (Kapitel 4.1) wird „peripheral data" auf theoretischer und methodischer Ebene beschrieben und mit den Methoden des Screen-Recordings, Maus/Tastatur-Trackings sowie der Logfiles hinsichtlich der Vor- und Nachteile verglichen. Nachteile bestehender Methoden sind etwa die notwendige manuelle Kodierung von Screen-Recording, die Abhängigkeit von installierter Software bei Maus/Tastatur-Tracking, oder die geringe Granularität von Log-Files. Auf technischer Ebene wird außerdem die Entwicklung eines Software-Frameworks namens „ScreenAlytics" und dessen Features zur Erfassung und

Analyse von peripheral data dokumentiert. Zusammenfassend können Forschende, nicht nur aus dem Bereich der pädagogischen Psychologie, die Software in bestehende Webseiten installieren und detaillierte Daten zur Interaktionen zwischen Nutzern und webbasierten Umgebungen aufzeichnen. ScreenAlytics verwendet diese Daten, um videoähnliche Replays und andere Visualisierungen wie z.B. Heat maps der Mausaktivität oder Navigationsgraphen zu erstellen. Replays können zudem mit Labels über Verhalten für qualitative Analysen versehen werden. Texteingaben können mit Metadaten zum Schreibprozess extrahiert werden. Eine Schnittstelle (API) ermöglicht es Forschenden, die gesammelten Daten in Echtzeit zu exportieren und benutzerdefinierte Ereignisse aufzuzeichnen. Zudem wird das System evaluiert und es werden Anwendungsszenarien diskutiert.

Das entwickelte Software-Framework bildet die methodologische Basis für die anschließend beschriebenen empirischen Studien, in denen Zusammenhänge zwischen dieser Datenquelle und SRL-relevante Variablen (Lernerfolg, Motivation, Kognitive Belastung und Verwirrung) untersucht werden. Auch für die Interventionsstudie liefert es die Datengrundlage. In dieser Studie werden die Auswirkungen von Learning Dashboards und metakognitiven Prompts auf den Lernerfolg untersucht.

In der ersten empirischen Studie (Kapitel 4.2) wurde der Zusammenhang zwischen Tippverhalten und Lernerfolg sowie aktueller Motivation untersucht. Die Argumentation der Aufzeichnung und Analyse von Tippverhalten ist, dass der Schreibfluss die dahinterliegenden kognitiven Prozessen beobachtbar machen kann. Der Schwerpunkt der Analysen liegt auf verschiedenen Indizes wie beispielsweise der Länge oder Häufigkeit der Pause oder Korrekturen. In der Studie wird angenommen, dass Indizes höherer Schreibgeschwindigkeit deshalb positiv mit Lernerfolg korrelieren. Bezüglich Motivation nennen Rheinberg und Kollegen (2001) Aufgabenbearbeitungszeit und die Qualität der Aufgabenbearbeitung als potentielle Indikatoren – diese Studie nimmt daher an, dass beide auch in Indizes des

Schreibprozesses Ausdruck finden. In einer Online-Lernumgebung zum Erwerb von deklarativem und prozeduralen Wissen über Website-Programmierung wurden $N = 43$ Studierende untersucht. Im Verlauf der Studie sollten Lernenden zunächst einen Beispielsatzes abschreiben um eine Baseline des Tippverhaltens erzeugt. Anschließend wurden initiale Motivation mit dem Questionnaire of Current Motivation (QCM, Rheinberg, Vollmeyer & Burns, 2001) und räumliches Vorstellungsvermögen mit dem VZ-2 Paper-Folding Test (Ekstrom, French, Harman & Dermen, 1976) als mögliche Einflussvariablen erhoben sowie das deklarative und prozedurale Vorwissen erfasst. Nachdem die Lernenden die Hälfte der Lernumgebung bearbeitet hatten, sollten sie ihr bisheriges Wissen zu drei Konzepten in eigenen Worten in einer Erinnerungsaufgabe aufschreiben. Zudem sollten sie zwei interaktive Aufgaben durch Schreiben von Programmiercode lösen. Nach diesen Aufgaben wurde jeweils eine Kurzversion des QCM präsentiert, um die aktuelle Motivation in hoher zeitlicher Nähe zum Tippverhalten zu erfassen. Nach dem Lernen füllten die Lernenden denselben Wissenstests aus, der vor dem Lernen bearbeitet wurde. Entgegen der Hypothese zeigte sich, dass Indizes niedrigerer Schreibgeschwindigkeit während der Erinnerungsaufgabe sowohl mit höheren Leistungen bei der Erinnerungsaufgabe, als auch mit höheren Werten im Lernerfolgstest korrelieren. Gleichwohl konnte gezeigt werden, dass Indizes höherer Schreibgeschwindigkeit während den interaktiven Programmieraufgaben mit höherem deklarativem und prozeduralen Wissenserwerb einhergeht. Dieses Muster fand sich auch bezüglich der aktuellen Motivation. Die Befunde des Zusammenhangs werden als Ausdruck von Aufgabenspezifität diskutiert. Beim Programmieren wird das Schreiben einer korrekten Abfolge vorher erlernter Code-Teile verlangt, weshalb Lernende mit schnellem und korrektem Abruf auch schnelleres Tippverhalten zeigen. Im Gegensatz dazu benötigt das Schreiben von Fließtext die Rekonstruktion und Verbalisierung von Wissen. Langsames Schreiben oder häufige Korrekturen werden eher als Ausdruck hoher Standards interpretiert.

In der Diskussion werden zudem methodologische Probleme besprochen und ausstehende Fragestellungen erörtert.

In der zweiten Studie dieser Arbeit (Kapitel 4.3) wurde experimentell untersucht, ob ein Zusammenhang zwischen den Mausbewegungen von Lernenden und deren kognitiver Belastung sowie affektiven Zuständen besteht. Dabei wird Mausbewegung als eine natürlich auftretende, sekundäre Aufgabe im Sinne des Dual-Task Paradigmas betrachtet. Der Grundgedanke ist, dass Pausen in der Mausbewegung entstehen, wenn die Belastung durch die primäre Lernaufgabe hoch ist. In einer quasi-experimentellen Studie wurden dafür $N = 49$ Studierende untersucht, die online zum Thema Website-Programmierung gelernt haben. Kognitive Belastung wurde über die Reaktionszeit gemessen. Lernende mussten bei einem Wechsel der Hintergrundfarbe der Lernumgebung so schnell wie möglich eine Taste drücken. Zudem wurde deklaratives und prozedurales Wissen sowie positiver und negativer Affekt mit dem PANAS-Instrument (Krohne, Egloff, Kohlmann & Tausch, 1996) jeweils vor und nach dem Lernen erfasst. In der Experimentalgruppe ($N = 28$) wurde die Messung der kognitiven Belastung nur dann ausgelöst, wenn über 6 Sekunden keine Maus- und Tastaturbefehle registriert wurden. In der Kontrollgruppe ($N = 21$) wurde die Messung in zufälligen Intervallen zwischen 15 und 35 Sekunden ausgelöst. Wie in der Hypothese vermutet, zeigte sich höhere kognitive Belastung in der Experimentalgruppe mit einem mittleren Effekt ($d = .60$). Zudem konnten signifikante Korrelationen zwischen Mausbewegungen und Affekt für die Kontrollgruppe gezeigt werden. Die Ergebnisse sind äußerst vielversprechend für eine Echtzeit-Messung dieser schwer zu erfassenden Variablen. Die Ergebnisse, einige methodische Limitationen der Studie sowie mögliche Anwendungen für Interventionen und weitere Forschungsarbeiten werden abschließend diskutiert.

In der dritten Studie dieser Arbeit (Kapitel 4.4) wurde untersucht, ob Verwirrung, subjektive und objektive Schwierigkeit von Items, sowie metakognitive Einschätzungen zum eigenen

Wissen durch Mausverhalten gemessen werden kann. Dafür wurden das Mausverhalten von N = 144 Personen bei der Beantwortung von Multi-Item Skalen aufgezeichnet. Multi-Item Skalen wurden gewählt, weil sie eine strikte Struktur aus Frage und zugehörigen Antwortoptionen einhalten, aber dennoch relevant für Lernumgebungen sind. Es wurden metakognitive Feeling-of-Knowing (FOK) Urteile zu 18 Items eines Tests zur kristallinen Intelligenz (BEFKI) abgefragt. Danach wurden 60 Items eines Big-Five-Inventory erfasst, von denen 6 mit falscher Grammatik oder Widersprüchen manipuliert waren um Verwirrung zu induzieren. Abschließend wurden die eigentlichen Antworten zu den BEFKI-Fragen abgefragt. Es konnte gezeigt werden, dass 1) manipulierte Items an erhöhten Indizes des Mausverhaltens erkannt werden können, 2) die Stärke der Manipulation (Widerspruch > Grammatik) anhand des Mausverhaltens erkannt werden kann, 3) höhere Indizes des Mausverhalten zwar mit höherer subjektiver Schwierigkeit von Items einhergehen, die Stärke dieses Zusammenhangs aber nicht zur Vorhersage der subjektiven Schwierigkeit ausreicht, 4) höhere Indizes des Mausverhaltens mit höherer objektiver Schwierigkeit einhergehen, aber niedrig korrelieren, 5) Fragen mit höheren FOK Urteilen längere Antwortzeiten aufweisen. Zudem wurden detaillierte Analysen zu verschiedenen Indizes des Mausverhaltens angestellt. Die Studie diskutiert die Ergebnisse unter Berücksichtigung bestehender Evidenzen im Bereich der Survey-Forschung. Zudem werden Limitationen der Studie und mögliche Anwendungen der Ergebnisse, unter anderem in Rapid-Assessment Tasks besprochen.

Die vierte und letzte Studie dieser Arbeit (Kapitel 4.5) untersucht die Effekte von Learning Dashboards zur Unterstützung des SRL sowie die detaillierte Interaktion von Lernenden mit dieser Intervention. Learning-Dashboards enthalten Visualisierungen von Daten über Lernprozesse, die zuvor gesammelt, verarbeitet und analysiert wurden. Der Grundgedanke der Studie ist, dass Learning Dashboards unterstützen, indem Informationen zum Lernprozess bereitgestellt werden und zusätzliche metakognitive Prompts dazu führen, dass diese

Informationen auch sinnvoll in Lernstrategien genutzt werden. Die Studie implementiert Empfehlungen aus bisherigen Reviews. So wurde berücksichtigt, 1) die Intervention selbst und die verwendeten Datenkanäle stärker theoretisch auf SRL-Frameworks zu fundieren, 2) nicht nur das Bewusstsein über den eigenen Lernprozess zu schärfen, sondern Lernende auch zu Veränderungen am Lernprozess zu bewegen, sowie 3) systematisch-experimentelle Designs zur Untersuchung der Effekte von Dashboards auf den Lernerfolg anzuwenden. Die Faktoren Prompt und Dashboard wurden dafür experimentell variiert. $N = 138$ Lernenden wurden dafür zufällig auf eine Kontrollgruppe ohne Intervention, eine Gruppe mit lediglich Prompts, eine Gruppe mit Dashboards sowie eine Gruppe mit Prompts und Dashboards verteilt. Inhalte waren die Grundlagen zur Programmierung von JavaScript in einer Online-Lernumgebung. Lernenden wurden zunächst kurze Videotrainings zur Nutzung der Interventionen gezeigt, anschließend wurde Einstellung zu Privatsphäre und metakognitives Strategiewissen als Kovariate erhoben. Auf einen deklarativen und prozeduralen Wissenstest folgte eine 60-minütige Lernphase. Nach 20 und 40 Minuten wurde die jeweilige Intervention präsentiert. Anschließend folgte derselbe Wissenstest sowie eine Evaluation des Dashboards und ein Selbstbericht zur kognitiven Belastung. Es zeigten sich keine signifikanten Unterschiede zwischen den Gruppen hinsichtlich des Lernerfolgs. Als Grund dafür wird hauptsächlich der fehlende Bedarf an Regulation wegen zu hoher vorgegebener Struktur der Lernumgebung genannt. Die detaillierte Nutzung der Interventionen sowie die entstehende kognitive Belastung und die wahrgenommene Nützlichkeit unterschiedlicher Teile des Dashboards werden ausführlich besprochen. Es werden außerdem Empfehlungen für weitere Forschung ausgesprochen.

Im Anschluss der Präsentation der fünf Forschungs- und Entwicklungsarbeiten werden diese im Gesamtzusammenhang gesetzt und der Einsatz von peripheral data in technologie-gestützten Lernumgebung kritisch reflektiert.

Danksagung (Acknowledgement)

Prof. Dr. Maria Bannert

Danke, dass Du meine Dissertation und alle damit verbundenen Studien, Anträge, Konferenzen und Hürden mit Wissen, Erfahrungen, Diskussionen, Ratschlägen und unerschütterlichem Vertrauen in meine Vorhaben und meine Arbeitsweise begleitet und unterstützt hast. Ich danke dir auch für die vielen lebensnahen Gespräche und Deine humorvolle, motivierende, selbstlose und immer faire Art. Nicht zuletzt möchte ich Dir danken, dass Du mich so selbstverständlich in Kontakt mit Deinen inspirierenden Kolleginnen und Kollegen aus der ganzen Welt gebracht hast.

Prof. Dr. Tina Seufert

Danke, dass Du mein Promotionsvorhaben bereits von Anfang an mit hilfreichen Rückmeldungen zu meinem Antrag, meinen Studien und Entwürfen unterstützt und begleitet hast. Danke auch für die herzlichen Begegnungen und Gelegenheiten, meine Forschung mit Dir und Deinem Team in unseren Kolloquien zu diskutieren.

Meine Kollegen an der Universität Würzburg

Danke, dass ihr mir den Einstieg in die Dissertation und die Universität als Arbeitsplatz so schön und den Abschied aus Würzburg so schwer gemacht habt.

Meine Kollegen an der TUM School of Education

Danke, dass ihr mir trotz der großen Distanz ermöglicht habt, ein Teil des Teams zu sein.

Meine internationalen Kollegen

Danke für die zahlreichen inspirierenden, formellen und informellen Treffen und Gespräche auf Konferenzen und E-CIR Meetings.

Meine Eltern und Brüder

Unter 100 deutschen Grundschulkindern werden im Mittel elf promoviert. Nur eines davon stammt aus Nichtakademiker-Haushalten (vgl. Stifterverband für die Deutsche Wissenschaft e.V., 2017, S. 12). Danke, dass diese strukturellen Ungleichheiten durch Eure Unterstützung in unserer Familie gleich dreimal überwunden werden konnten.

Meine Freunde

Ihr sorgt für die Raison d'Être und das Fundament allen Tuns. Danke, dass ihr immer da seid!

Cusanuswerk, Studienstiftung des deutschen Volkes, Deutsche Gesellschaft für Online-Forschung und deren Mitarbeiter und Mitarbeiterinnen

Danke für die finanzielle und ideelle Förderung. Danke insbesondere für die Bearbeitung (und die Bewilligung!) meiner zahlreichen Anträge für Konferenzen, Weiterbildung und Studien.

Table of Content

1 Learning in Technology Enhanced Environments: A Glance at Self-Regulation

The rapid development in technology during the last decades led to an intensive use of technologies for learning in almost all formal and non-formal educational settings, starting with basic offline computer applications in the late 1970s and reaching sophisticated online learning environments including simulations, intelligent agents, and virtual or augmented reality nowadays (e.g., Harting & Erthal, 2005; Martín-Gutiérrez, Mora, Añorbe-Díaz & González-Marrero, 2017). Since almost two decades, the number of US-students taking online courses consistently grows and more than 28% of higher education students are enrolled in at least one online course (Seaman, Allen & Seaman, 2018). Moreover, formal and non-formal massive open online courses (MOOCs) continue to grow in both the number of offered courses and the volume of learners enrolling (Shah, 2015).

These ongoing developments brought overwhelming, unprecedented possibilities, and led to an ubiquitous availability of a constantly growing, inconceivable amount of information. Thus, digital media has many inherent advantages over non-digital for learners, such as location-independent access to study materials, more interactive contents, or multiple sources and perspectives to choose from for a topic learners want to study.

At the same time, besides the euphoric expectations that we have on digital media, some of the challenges that learners experience did not change as they are independent of the media that is used to present content. For example, understanding the main ideas of a text does not differ just by changing the media from printed to digital - although digital native readers have a preference to read with digital devices (e.g., Singer & Alexander, 2017). It's rather a matter of what pedagogical role a new medium is able to take than through what medium a content is presented. Other challenges even occurred only as a consequence of the possibilities that digital media brought. For example, hypermedia provides non-linear navigation which (mostly) is not available in non-digital media and thus, requires students to additionally search for hyperlinks and judge whether these are relevant for their learning goals (Bannert & Mengelkamp, 2013).

One crucial point in learning with technology enhanced learning, is the often low (or even missing) external guidance compared to traditional educational settings where lecturers, teachers and peers provide regulation for the learning process. This means that learners have to take care of activities like goal-setting, planning the steps to achieve these learning goals, monitoring the progress, and selecting appropriate learning strategies - they have to regulate their often dynamic and complex learning processes themselves, an activity that is referred to as self-regulated learning (SRL) and that has been focused in educational psychology during the last decades (e.g., Winne & Nesbit, 2009; Zimmerman, 2008). In an early definition, self-regulating students are described as “metacognitively, motivationally, and behaviorally active participants in their own learning process” (Zimmerman, 1986, p. 308). Within this definition, “metacognitive” refers to the planning, monitoring, organization, and evaluation of one’s own learning, “motivational” refers to perceive oneself as competent, self-efficacious, and autonomous, and “behavioral” refers to selecting, structuring and creating conditions that are best suitable for their learning (Zimmerman, 1986, 2000).

SRL empowers learners to independently acquire new skills and knowledge, and SRL competencies are an important predictor for educational and academic success (Dent & Koenka, 2016). Hence, both researchers (Dignath, Buettner & Langfeldt, 2008) and policy makers (Pirrie & Thoutenhoofd, 2013) argue that successful SRL is a key competence to successfully cope with the dramatically fast changes of a modern, knowledge-based society.

However, digital media does not only require SRL but, compared to non-interactive traditional materials like books or videos, the technological achievements also provide promising new ways to support SRL – both during the actual learning process, and even in the long term to improve general SRL skills beyond a single intervention for the current learning process. In order to achieve a development of SRL in learners, Zimmerman (2001) emphasizes that opportunities have to be provided for learners to practice SRL strategies. A prominent example for such interventions are metacognitive prompts (Bannert, 2007, 2009; Berthold, Nückles & Renkl, 2007; Nückles, Hübner & Renkl, 2009). Based on the finding that learners who actually possess metacognitive strategies often have difficulties with applying appropriate metacognitive activities while learning (so-called production deficit, e.g., Bannert & Mengelkamp, 2013; Veenman, Van Hout-Wolters & Afflerbach, 2006), metacognitive prompts aim at triggering them to achieve better learning outcomes. Metacognitive prompts can be presented to the learner in different modalities, from low-level cues that just present self-directed questions or instructions to a more sophisticated delivery through pedagogical agents or intelligent tutoring systems (Azevedo et al., 2012; Azevedo, Johnson, Chauncey & Burkett, 2010). Another example for an intervention, that this work will look at, are learning dashboards, that aim at support metacognitive activities by informing the learner about their current learning process through presenting visualizations of different aggregated indicators (e.g., Schwendimann et al., 2017; Teasley, 2017).

Although research shows that such interventions are effective instruments to support learning, they are mostly designed as a “one-size-fits-all” intervention, meaning the same interventions are presented to all learners, regardless of their prerequisites. At the same time, a range of studies in the area of instructional design find aptitude-treatment interaction(ATI) effects indicating that characteristics of learners such as prior knowledge moderate the effects of interventions (e.g., Seufert, 2003). Thus, aiming at higher effects on learning outcomes, interventions can be designed that are adaptive and successfully address learners’ diverse prerequisites (e.g., Shute & Zapata-Rivera, 2008). To do so, one needs to know and decide what variables are relevant for the instructional milieu that should be supported (e.g., learner characteristics like prior knowledge or motivational state), and then acquire valid measures of these variables from available data sources (e.g., Vandewaetere, Desmet & Clarebout, 2011), which are fed into the adaptive target element in order to support and enhance learning.

Models of SRL (e.g., Boekaerts, 2007; Winne & Hadwin, 1998; Zimmerman, 2000) provide suggestions for variables that are important for the learning process and that therefore should be taken into account for adaptions on interventions (e.g., motivation, affect, metacognitive knowledge). Regarding an accurate diagnosis (i.e., the acquisition of valid data on variables relevant for learning), self-reports and multi-item scales that learners fill prior to learning are still a common methodology but face several drawbacks such as being subjective, obtrusive or reactive to the measure, disturbing the learning process, not being available during the learning process, and not being able to capture the high granularity of adaptions that learners make (e.g., Zimmerman, 2008). These disadvantages and solutions to it are currently under debate in research on technology-enhanced education. There is a crucial need for more objective, non-obtrusive, real-time data sources both for a better understanding and verification of theories of learning with a focus on the recurring processes that occur, and for adaptive learning systems. Winne and Perry (2000) emphasized the need for “on-the-fly” and “online” measures especially for SRL. This need led to an extension of the methodological repertoire in the research on educational psychology and technology enhanced learning that provide a range of different data streams. Examples are data streams like eye tracking (e.g., used in Miller, 2015), psychophysiological measures (e.g., EDA, EKG, EEG, McQuiggan, Mott & Lester, 2008), camera-based recognition of facial expressions (Baltrusaitis, Robinson & Morency, 2016), concurrent think-aloud (Bannert & Mengelkamp, 2008; Greene, Robertson & Costa, 2011), or web log files (Cocea & Weibelzahl, 2006). Recording these data channels became relatively straightforward and affordable. However, using appropriate data sources is just a requirement of the subsequent challenge to find indices that can contribute to measure important features of the learning process - the data still needs to be processed, analyzed and interpreted regarding variables that are relevant for learning. Most data channels can only act as a proxy for learning behavior and research is needed that uncovers relationships between patterns in data channels and variables of interest. For example, EDA signals can easily be recorded, but systematic, rigorous controlled studies need to show how these signal correspond with variables of interest (e.g., Pijeira-Díaz, Drachsler, Järvelä & Kirschner, 2016). This is typically done by 1) identifying externally observable behaviors in the data channel (e.g., EDA signal peaks), 2) identifying latent states (e.g., regulatory activity after the peak) that are linked to these observable behaviors and 3) discovering patterns in the latent states that explain variance in the learning outcome (these steps are adapted from Reimann, Markauskaite & Bannert, 2014 who described them for sequence mining). Thus, gathering valid interpretations and inferences regarding the learning process from collected data is still very challenging.

A data source that has hardly been discussed in this discourse on examining learning in technology-enhanced environments is so-called peripheral data, that is addressed in this work. Like traditional log files, peripheral data represents the interaction between learners and online environments as a chronological sequence. However, peripheral data has a very high granularity. Instead of simple page statistics, detailed events of mouse, touch, and keyboard input devices as well as the website contents are recorded with a high frequency that later enables us to reconstruct the complete observable interaction as a simulated replay similar to a screen recording. Compared to other methods like screen recording or log files, peripheral data has some important advantages: Peripheral data opens the black box of classic log files that only gives insight into which page was accessed when, but not what actually happened on that page. Most importantly, it keeps the acquired data automatically processable as it is not represented as a pixel-based video file. Moreover, peripheral data is available in real time, needs no manual coding of events and has no software or hardware dependencies on the learner's computer.

This work contributes on closing an existing research gap in the described challenges of using peripheral data in technology-enhanced learning on multiple levels. First, on a theoretical and methodological level, an approach to record and analyze detailed, event-based peripheral data is described and a software framework called ScreenAlytics was developed, that enables researchers to easily acquire that data in their studies. Secondly, on an empirical level, studies of this work investigated the correspondence between peripheral data and variables relevant for (self-regulated) learning (i.e., cognitive load, affect, motivation and confusion). Thirdly, on an intervention level, peripheral data was used as a real-time input source for learners to inform them about their own learning process (i.e., learning dashboard). This dashboard was empirically examined regarding its impact on the learning outcome and the detailed usage of such an intervention

2 Theoretical Background

This chapter introduces general theories and assumptions that are needed to understand the research and development of this work. Note that it only reviews theoretical concepts and constructs that are relevant to all presented studies, i.e., cognitive load (CL), SRL, and metacognition. Constructs that are solely related to one specific study are described in the theory chapter of the according study, i.e., affect, motivation and confusion. The chapter starts with CL, introduces the basic idea of SRL and important models of it, and continues with presenting metacognition as a construct closely related to SRL. After that, the idea and the current state of how SRL processes are measured using multimodal data streams is briefly introduced. Finally, instructional interventions are addressed that can support learners in regulating their learning, i.e., adaptive learning systems, prompts, pedagogical agents, learning dashboards, and eye-movement modeling examples.

2.1 Cognitive Load in Multimedia Learning Environments

Multimedia learning environments are characterized by the representation of content in different formats. Following the basic assumption of multimedia learning, people learn better from text and image (multiple representations) than from text alone (Mayer, 2009; Schnotz, Seufert & Bannert, 2001). The integration of information from different formats enables learners to construct an elaborate mental model about the facts to be learned, which constitutes "understanding" and allows transfer (e.g., Mayer, 2009; Schnotz & Bannert, 2003). A prerequisite to integrate information is the processing and transfer of information acquired by our sensory organs from the sensory memory to the conscious working memory. From there, information can be stored in the long term memory, from where it can be recalled again into the working memory (e.g., Atkinson & Shiffrin, 1971). Hence, working and long term memory are central, interacting cognitive structures (Sweller, 2005). Cognitive load theory (CLT), developed by Sweller (1988, 2005), states that learning is therefore always connected to CL in the working memory. CLT provides one of the most important frameworks for research on learning and instruction. Moreover, and maybe even more importantly, it also provides guidelines on the efficient design of learning environments (Plass, Moreno & Brünken, 2010; Sweller, Ayres & Kalyuga, 2011). CLT has been confirmed in a whole range of empirical studies and reviews (e.g., Sweller, 1994, 2004, 2005; Sweller & Chandler, 1994; Sweller, Van Merrienboer & Paas, 1998; van Merriënboer & Sweller, 2005). A fundamental claim of the theory is that the working memory is limited by two factors: the number of information and the duration that one can keep information in the working memory (e.g., Baddeley, 1992; Sweller, 2009). George Miller (1956) already suggested the number 7 (plus/minus 2) as the "magic number" that can be kept in the working memory by human beings and that characterized the memory limit. Later, researchers revised this number to 2 to 4 elements that can be kept in the working memory simultaneously (Cowan, 2000; Sweller, 2004). Regarding time, the working memory is able to store information for a maximum of 20 to 30 seconds (e.g., Kirschner, Sweller & Clark, 2006). These limitations have to be considered in the design of learning materials according to CLT. CLT claims that learning is reduced if the processing demands of the learning task exceeds this capacity of the working memory – learners experience a so-called *cognitive overload* (Mayer & Moreno, 2003).

Total experienced CL consists of three additive components (Moreno & Park, 2010; Sweller et al., 1998). Intrinsic cognitive load (ICL) refers to the structure and complexity of learning materials (Sweller & Chandler, 1994) and is characterized by the level of content interactivity. This level of interactivity depends on the amount of interrelated information units that have to be kept in the working memory to understand the learning material (Brünken, Steinbacher, Plass & Leutner, 2002). If many elements are simultaneously needed in the working memory (e.g., when learning how different parts of a motor interact), the ICL is high. If elements can be processed consecutively (e.g., when studying vocabulary), ICL is low. Moreover, ICL depends on the prior knowledge of learners related to the learning content. The higher the prior knowledge, the lower the intrinsic load induced by the learning materials (Sweller, 1994, 2005).

Extraneous cognitive load (ECL) is related to the way in which materials are presented (Sweller & Chandler, 1994). The more cognitive resources a learner needs to extract information from the presented materials, the higher the ECL. As ECL does not contribute to learning, but is needed only to extract information (Brünken, Plass & Leutner, 2003), it should be kept low by proper instructional design (e.g., Bannert, 2002).

Germane cognitive load (GCL) is the third source and describes the cognitive effort needed for constructing and automating schemata in the long-term memory (Sweller, 2005). The concept of schemas has been described by Piaget (1928) and Bartlett (1932). Schemata organize the storage in the long-term memory and make information available efficiently. Hence, a high GCL represents efficient learning. However, this type of load is debated in literature as a potential circular reasoning is criticized (GCL is high, learning is better; learning is better, GCL is higher, e.g., Kalyuga, 2011), and the differentiated measurement of single loads in general, but especially of GCL is not straightforward (Gerjets, Scheiter & Cierniak, 2009; Kirschner, 2002; Klepsch, Schmitz & Seufert, 2017; Schnotz & Kürschner, 2007; van Gog & Paas, 2008). Later literature on CLT also distinguish between productive load, including intrinsic and germane load, opposed to unproductive load, which is extraneous load (e.g., Paas & Ayres, 2014).

Moreover, Seufert (2018) most recently explained how CLT and SRL (which is introduced in the next chapter 2.2) are conceptually related – a connection that has been neglected for the most time during the largely separate development of both theories. She argues that self-regulation is a highly demanding process, because learners do not only need to handle the actual learning task, but also need to invest cognitive and metacognitive resources in all phases of SRL, such as monitoring or goal-setting. By this, SRL causes ICL to the learner. However, it is worth noting that according to her model, regulation can also cause unproductive load through regulatory activities or off-task demands that disturb the learning process. In her model, she argues that the difficulty of a task determines the imposed load and hence, the free resources depend on this task difficulty as well as the individual capacity of learners. Only if there are enough free resources, regulation is possible at all. However, that does not mean that regulation increases linearly with more resources being available. As easy tasks might not need regulation while difficult tasks may not allow for regulation because it allocates too many resources, an inverse-U shaped relation between task difficulty and regulatory activities is described.

2.2 Self-Regulated Learning Frameworks

Since almost three decades, SRL has been (and still is) an important field that gained immense attention in educational research and widely influenced educational practitioners. This is not surprising, as learners' ability to steer their learning processes is considered as highly important, especially in a knowledge society (e.g., Azevedo & Greene, 2010). Moreover, constructs and frameworks of SRL integrate (meta-)cognitive, motivational / affective, social and behavioral components of theory and research (Boekaerts & Niemivirta, 2000, p. XXII).

However, reconciling so many facets of learning also led to a lack of consistency in definitions and operationalizations and in consequence, a lack of congruency in theory and empirical knowledge. Hence, there seems to be no straightforward or simple definition of SRL (Boekaerts & Corno, 2005). Rather, constitutions, processes, aims and challenges in the scope of learners' self-regulatory activity can be described.

Self-regulation refers to “self-generated thoughts, feelings and actions that are planned and cyclically adapted to the attainment of personal goals” (Zimmerman, 2000, p. 14). Theories of self-regulation were developed and used not only in the context of learning, but are of high relevance also in other disciplines like clinical or organizational psychology. In the context of learning, regulatory activity is important as learning is a complex and dynamic process that needs to be planned well, and that includes a range of states that need to be monitored and controlled. Regulation during learning does not have a single source but is rather fed by a continuum from internal (i.e., learners themselves) to external sources, which can be lecturers, teachers, peers or even computer programs. However, even if there is external regulation, learners need to self-regulate parts of their learning (Boekaerts & Corno, 2005). Learners that successfully regulate their learning are described as “metacognitively, motivationally, and behaviorally active participants in their own learning process” (Zimmerman, 1986, p. 308).

Hadwin, Järvelä and Miller (2017) tried to define fundamental constitutions of regulated learning (although not only limited to, these are relevant for *self*-regulated learning) and argue that SRL always 1) is intentional and goal-directed, 2) involves metacognitive planning, monitoring and control, 3) involves regulation of behavior, cognition, and/or motivation/affect, but is not about the construction of domain knowledge, 4) depends on the social surround and/or interplay and 5) requires opportunities (challenges) to apply regulatory activity.

Within the last decades, a range of theoretical frameworks were developed. There exist excellent reviews of these models, Puustinen and Pulkkinen (2001) reviewed five models: Boekaerts and Niemivirta (2000), Borkowski (1996), Pintrich (2000), Winne and Hadwin (1998), and Zimmerman (2000). More recently, Panadero (2017) conducted a partly intersecting review of six SRL models: Boekaerts and Corno (2005), Efklides (2011), Hadwin, Järvelä and Miller (2017), Pintrich (2000), Winne and Hadwin (1998), and Zimmerman (2000).

These reviews present the models, evaluate its empirical validation and compare them with each other. Thus, this work will not review the models in detail again, but rather concentrates on relating the measures and intervention that were used in the empirical studies and the software development of this work on existing models. However, as they represent different views and help to understand the historical development, three of the models are quickly introduced: Zimmerman's triadic model (Zimmerman, 1989), Zimmerman and Moylan's cyclical phases model (Zimmerman & Moylan, 2009), and Winne and Hadwin's COPES model (Winne & Hadwin, 1998).

In his Triadic Analysis of SRL (Figure 1), Zimmerman (1989) describes how SRL can be implemented in Bandura's triadic model of social-cognition (Bandura, 1986). The model represents the interactions of three SRL determinants, namely the environment, the behavior and the person (self-)level. The core idea is that SRL is not solely determined by individual processes, but influenced by environmental and behavioral events. Moreover, the relationships between determinants are reciprocal, but not necessarily symmetric in strength and temporal patterning. As an example, using a self-evaluation strategy such as checking the math homework (behavior level) will provide information on accuracy and whether checking needs to be continued through enactive feedback (from behavior to person). An example for environmental influence can be the arrangement of a quiet study area which involves proactive behavior such as eliminating noise or changing light conditions. The continued use of this setting depends on the effectiveness indicated reciprocally through the environmental feedback loop. It is important that, in order to be labelled as self-regulated, learning strategies need to be triggered from key personal processes (such as goal-setting), and not from external instruction. Zimmerman (1989) argued that individual's covert processes (e.g., an elaboration strategy) in the model also reciprocally affect each other and already mentions that metacognition plays as an important role within the covert feedback loop.

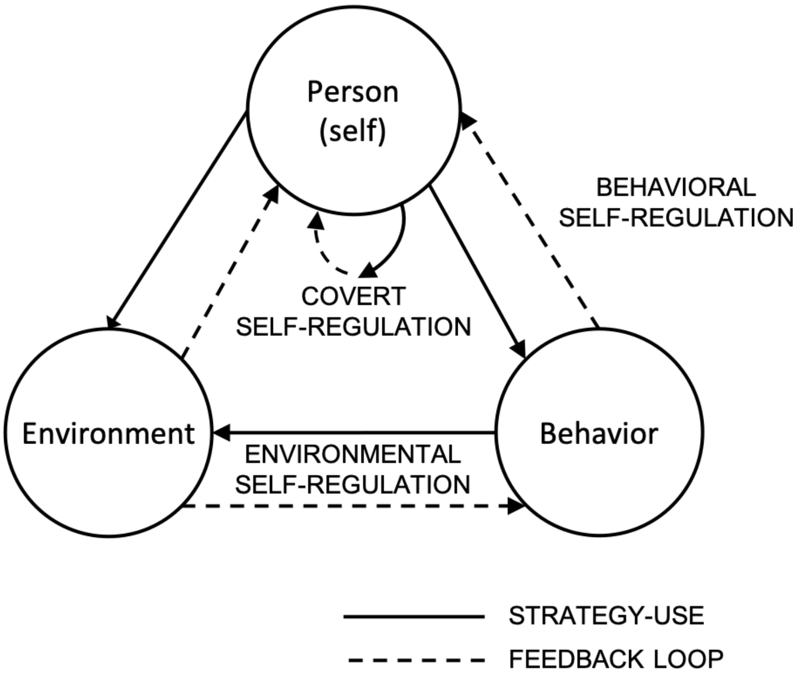


Figure 1. Triadic analysis of self-regulated functioning, adapted from Zimmerman (1989) including the updates of Zimmerman (2013)

An important difference between the cyclical phase model of Zimmerman and Moylan (2009) and the aforementioned triadic analysis model can already be recognized in its name. It distinguishes three phases in a recurring manner: forethought, performance and self-reflection, as shown in Figure 2. The model explains the interrelation of motivational and metacognitive processes at the person level, and as a process-oriented model, it also describes adjustments of the learning processes through learners using recurrent feedback-loops. In the forethought phase, learners analyze the task, set goals, plan ways to achieve them and motivational states trigger the learning process and activate learning strategies. During the performance phase, learners carry out the actual task, monitoring how they progress, and apply a range of self-control strategies to remain cognitively active and motivated in order to complete the task. In the self-reflection phase, students evaluate how they have performed in the task and make judgements about their success or failure. Self-reactions are generated by these attributions, which can have a positive or negative impact on how learners will tackle tasks in later performances. The presented version in Figure 2 is the latest available version of the model, although earlier versions were already presented in Zimmerman (2000), not including the subprocesses of the phases.

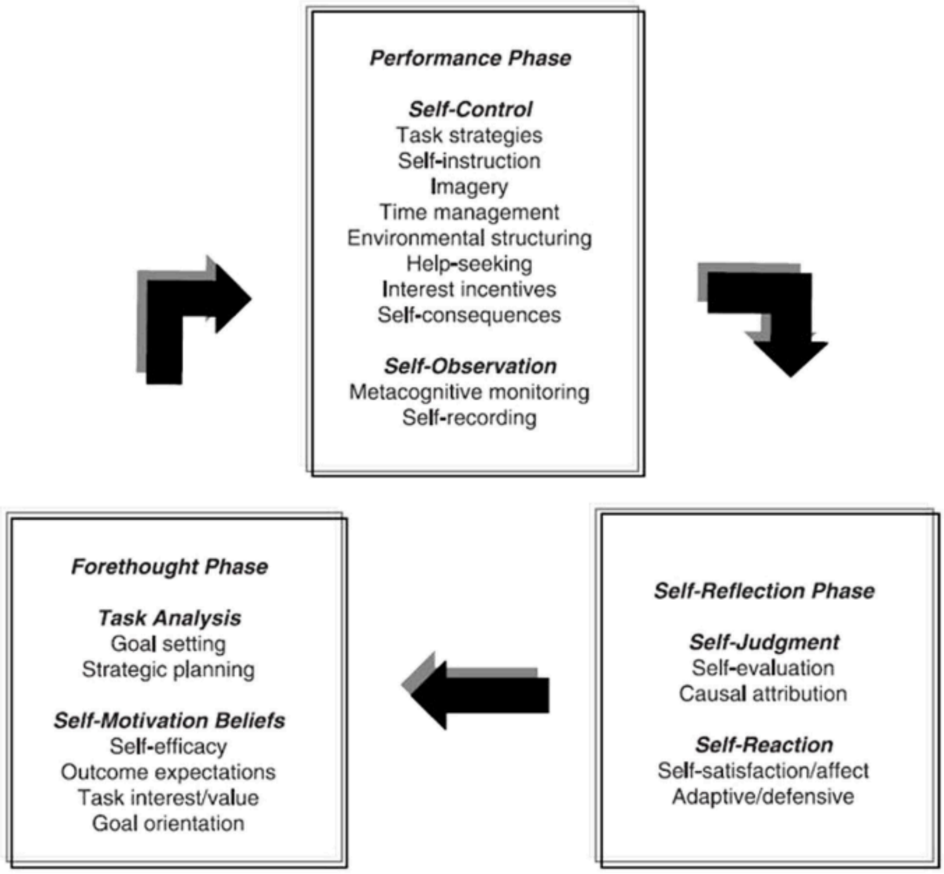


Figure 2. Current version of the cyclical phases model, adapted from Zimmerman and Moylan (2009).

As depicted in Figure 3, Winne and Hadwin (1998) propose that learning occurs in four basic, recursive phases: (1) task definition: learners get an understanding of the task to be performed, (2) goal-setting and planning: learners generate goals and plan how to achieve these, (3) enacting on studying tactics and strategies: actions are needed to reach the goals, and (4) adaptations to metacognition: once the main processes are completed, learners take decisions on long-term changes in their motivation, beliefs and future strategies.

In their model, each of the four phases is described regarding the interactions between a learner's conditions, operations, products, evaluations and standards, which builds the acronym and the name of the model: COPES (as explained in Greene & Azevedo, 2007; Winne & Hadwin, 1998).

Conditions describe available resources and constraints that are inherent to a task or the learning environment. The model distinguishes between cognitive conditions including internal prerequisites such as beliefs, dispositions, or domain knowledge, and task conditions including external factors such as available resources, or instructional cues. Operations represent the cognitive processes, tactics and strategies that are used by the learners, including searching, monitoring, assembling, rehearsing and translating (referred to as SMART by Winne, 2001). Products are resulting information, created through operations, e.g., new knowledge. Examples for such products can be the definition of a task in the first phase, or the ability to recall a specific information while applying strategies (phase 3). Evaluations result from learners monitoring on how the products deviate from set standards, either generated internally or provided through external sources such as a teacher or a peer. Hence, a low fit between products and standards can result in further applying studying tactics, revise the conditions or standards, or both. Standards represent the criteria which learners take as the desirable end state of any phase they are currently in. Each aspect of a learning task might have different criteria that a learner actively determines. The overall criteria that are set in the task definition phase build the standards and thus, the learners goal (Greene & Azevedo, 2007; Panadero, 2017; Winne & Hadwin, 1998). It is important to note in this model that it is a "recursive, weakly sequenced system" (Winne & Hadwin, 1998, p. 281). Within in this system, the monitoring of products and standards in one phase can update the products from previous phases (Greene & Azevedo, 2007).

This work looks at SRL from the theoretical perspective of Winne and Hadwin's COPES model (Winne & Hadwin, 1998) for several reasons: 1) it is an actively used SRL model to date (Panadero, 2017), 2) it has a strong emphasis on metacognition (Panadero, 2017; Puustinen & Pulkkinen, 2001), 3) it is widely used in technology-enhanced learning settings, 4) it is reflective of SRL in older students / adults who encounter more cognitively demanding tasks.

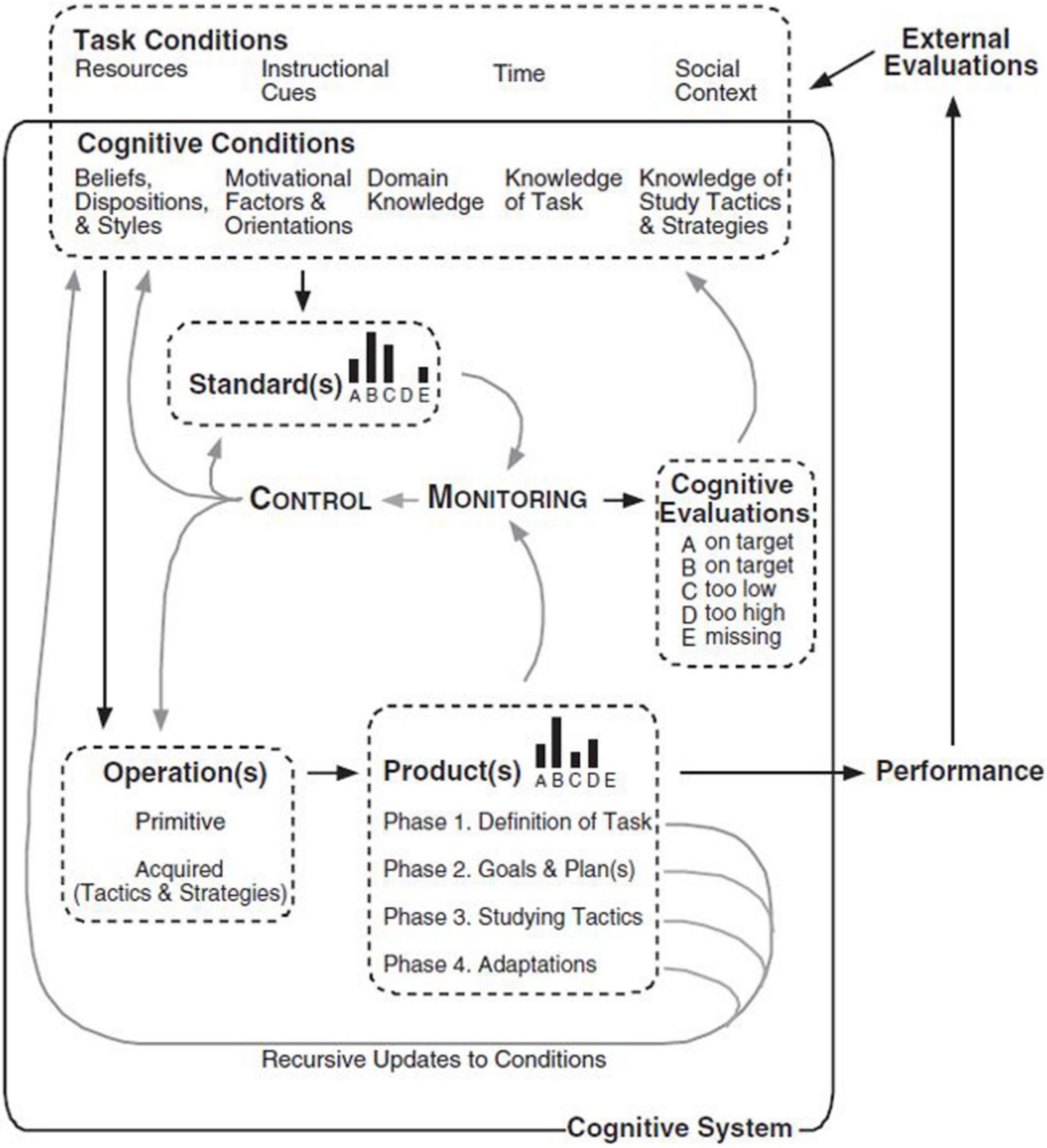


Figure 3. COPES model by Winne and Hadwin (1998).

2.3 Metacognition

Metacognition is a construct defined as thinking about one´s own thoughts and cognition in order to regulate one's own cognition (Flavell, 1979). Metacognition has two central components: monitoring and control (e.g., Nelson, 1990). Metacognitive monitoring "refers to the subjective assessment of one's own cognitive processes" (Koriat, Ma'ayan & Nussinson, 2006, p. 38). Monitoring processes therefore lead to a meta-level mental model of one's own cognition. As an example, learners compare a current state in their learning process with a target state (standard) and evaluate the achievement of a goal in order to update the mental model (Hadwin et al., 2017). On the other hand, control ”refers to the processes that regulate cognitive processes and behavior" (Koriat et al., 2006, p. 38). The discrepancy between achieved and desired states gives learners an opportunity for regulation in their learning processes. Thus, metacognition is a central construct in SRL (Winne & Hadwin, 1998).

Metacognitive knowledge can be distinguished from metacognitive skills (e.g., Hartman, 2001). Metacognitive knowledge describes knowledge that learners have about the interaction between tasks, person and characteristics of strategies (Flavell, 1979) while metacognitive skills are skills that learners have in order to apply metacognitive activities for controlling and monitoring their cognitive activities (Veenman, 2005).

Schraw (1998) distinguished knowledge of cognition from regulation of cognition. Knowledge of cognition is further specified in declarative, procedural and conditional knowledge. He argues that declarative metacognitive knowledge is knowledge about cognition, including general facts such as capacity limitations of the working memory, but also knowledge about the own cognition, such as individual conditions that influence one's learning process. Procedural metacognitive knowledge describes knowledge about actually enacting in the learning process, mostly represented as heuristics and strategies (e.g., chunking or categorizing information). Conditional knowledge refers to the when and why of using declarative and procedural knowledge. Regulation of cognition refers to “a set of activities that help students control their learning” (Schraw, 1998, p. 114).

Another aspect of metacognition are so called metacognitive judgments and feelings, that learners perform during their learning process to monitor their learning (Nelson, 1990). Different types of judgments and feelings can represent different aspects of monitoring: feelings of knowing (FOK), feelings of difficulty (FOD), judgments of knowing (JOK), judgments of learning (JOL), confidence judgments, etc. Efklides (2008) states that these results of metacognitive monitoring activate metacognitive skills. Learners feel / judge that there might be an issue in their learning process and use metacognitive skills to enact on it.

The discrepancy between achieved and desired states while monitoring gives learners an opportunity for regulation in their learning processes, which is one reason for metacognition being a central construct in SRL. Moreover, it has been shown that learners who use more metacognitive activities tend to show better learning outcomes (Veenman, 2005, 2011).

For both theory and empirical research, it is not trivial to unravel the mechanisms and characteristics of metacognition and cognition (e.g., Veenman et al., 2006), and there is only limited agreement on definition and terms of metacognition (Dinsmore, Alexander & Loughlin, 2008) as well as methods for the measurement of it. This led to the application of new and lavish multimodal data streams and methods such as think aloud (Bannert & Mengelkamp, 2008) and process mining (Sonnenberg & Bannert, 2016) of multimodal data streams in order to gain better insight into metacognitive processes during (self-regulated) learning, which will be further investigated in the next chapter. In this thesis, metacognitive activities are seen as a component of the broader theoretical construct of SRL. As described, these activities arise from learners’ metacognitive knowledge and skills.

2.4 Data Sources for Analyzing Self-Regulated Learning

As models of SRL describe processes that depend on and implement different latent constructs such as motivation, emotion, cognition and metacognition, and different aspects of it, it is not feasible to operationalize SRL as a whole. Rather, single components of SRL need to be measured and aligned. A combination of these measures can deliver insight into the interdependent phases of SRL. However, these components are also mostly not directly measurable but need to be inspected through operational definitions (e.g., Winne, Jamieson-Noel & Muis, 2002).

A commonly used and established operationalization for such latent constructs are self-reports. These are acquired prior, during or after learning either with questionnaires or in open formats using different modalities. Self-reports fulfil major methodological requirements for an accurate measurement, questionnaire instruments are usually tested and calibrated with extensive effort, and open self-reported formats are cross-validated by multiple raters. As such, they have provided the largely valid data basis for an enormous part of the findings in (educational) psychology. However, they also suffer from disadvantages, which are especially crucial in the dynamic context of SRL.

Firstly, self-report measure are of course, subjective. Main drawbacks of subjective measures are that 1) they suffer from systematic biases related to effects of order, scale, social desirability or memory (e.g., Bertrand & Mullainathan, 2001; Podsakoff, MacKenzie, Lee & Podsakoff, 2003), 2) correlations with objective measures of the same construct, if available, are found to be low for a range of constructs (Bommer, Johnson, Rich, Podsakoff & MacKenzie, 1995), and 3) they cause difficulties regarding the aggregation and interpretation because of their ordinal scaling (e.g., Sullivan & Artino, 2013).

Secondly, self-report measures are obtrusive. As such, they have the potential to disturb or interrupt the actual learning process and, hence, be a reactive measure that impacts the results for the measurement of learning outcomes or other variables of interest.

Thirdly, and most importantly, self-reported measures only provide a snapshot of the measured variable at a certain point in time. While this is not problematic for static learner characteristics, it is a major drawback for dynamically changing variables such as motivation, or affective and emotional states. Learners' self-reports are not capable of capturing the granularity of adaptions that they make during the learning process (e.g., Zimmerman, 2008). For both advancing theory and supporting learning, researchers and instructional designers would need such variables to be recorded "on-line", thereby reflecting changes in a continuous data stream that is available in real-time (e.g., Winne & Perry, 2000).

These drawbacks result in a demand for objective, unobtrusive, unreactive, continuous, online measures that has led to an extension of the methodological repertoire and data streams used in the research on educational psychology and technology enhanced learning. Among others, such process-related data sources include screen recordings, facial recognition data, eye tracking, video observations, log files, and physiological sensors (e.g., EDA, EMG, EEG, EKG, fMRI).

It is important to understand that process measures per se do, by no means, fulfil all mentioned demands and that the characteristics are independent from each other. For example, concurrent think aloud protocols depict the process, but are still a subjective measure (e.g., Sonnenberg & Bannert, 2018), retrospective think aloud protocols are continuous, but not available in real-time. Moreover, measures differ regarding their obtrusiveness and, as a consequence, in their reactiveness. For example, log files are unobtrusive, as learners may not even know that researchers (without privacy awareness) are capturing how they navigate through learning environments. As such, log files will not affect other measures. In contrast, EEG or EDA measures require devices to be attached on the learners and thus, are very obtrusive and might be reactive to other operationalized variables (e.g., emotional states or attention). Another aspect that differs between measures refers to the time resolution of process measures. While, for example, facial recognition of emotions might have a high frequency, changes in EDA signals that can be interpreted regarding CL have a lower time resolution (e.g., Setz et al., 2010), and coded behavior from video observations regarding SRL phases are available even less frequently (e.g., Järvelä, Volet & Järvenoja, 2010).

In many cases, the inferences that can be drawn from process data are ambiguous. For example, log files may indicate that there has been no interaction with the learning environment, but the missing interaction can have multiple reasons (e.g., a learner reads carefully a text in the learning environment without controlling it, or he/she is no longer sitting in front of the computer). In order to draw meaningful conclusions, it is often necessary to triangulate different data channels (resulting in so-called multi-channel data), that complement or validate each other (e.g., eye-tracking with facial expressions of emotions and screen-recordings).

Moreover, it is crucial to understand that most of these measures can only be used as proxy measures of relevant latent psychological variables (e.g., EDA for emotional states, Henriques, Paiva & Antunes, 2013) or need to be coded regarding a specific behavior (e.g., screen recordings or video observations, e.g., Malmberg et al., 2018). The coding of this behavior can either be done by researchers, or, this work can be assigned to the learners themselves in a subtle way as suggested in a paradigm used by the gStudy / nStudy software (Beaudoin & Winne, 2009; Hadwin, Nesbit, Jamieson-Noel, Code & Winne, 2007; Perry & Winne, 2006; Winne & Hadwin, 2013; Winne, Nesbit & Popowich, 2017). This software provides a toolkit for learners, each representing phases and levels of SRL. As an example, learners can set their goals using a specific type of note in the nStudy browser or tag information with descriptive, evaluative and action tags. This leads to log files that already are interpretable regarding SRL.

For other process data channels such as eye tracking or physiological devices, recording them has become rather easy from a technical perspective. However, it is very challenging to interpret them regarding learning processes from both theoretical and epistemic perspectives. In the context of process analyses on SRL, Reimann, Markauskaite, and Bannert (2014) characterize three steps for constructing theoretical explanations from recorded sequential event data. Although these data streams do not necessarily need to be sequential (e.g., sequences are not crucial when interpreting the mean number of gaze transitions in eye tracking data), the approach can be adapted and generalized on other data streams. Figure 4 illustrates important steps towards interpreting data with the example of how pauses in mouse interactions relate to CL – a question that is addressed in this work (see chapter 4.3).

Step 1: Identify externally observable behaviors

In a first step, externally observable behaviors have to be identified that can be represented as possible quantified indicators for latent psychological constructs. As an example, regarding mouse and keyboard data, every record consists of a timestamp and an event triggered by the learner (e.g., 15 seconds after beginning an exercise, the learner moved the mouse to position X/Y). As the raw data can be complex, potentially meaningful indices that describe behavior have to be extracted by aggregating, computing means, sums, and ratios. These generated indices (also referred to as features) represent a variety of observable information about learner behavior, e.g., number of clicks on a specific element, frequency of pauses, changing focus between elements in a learning environment, etc.

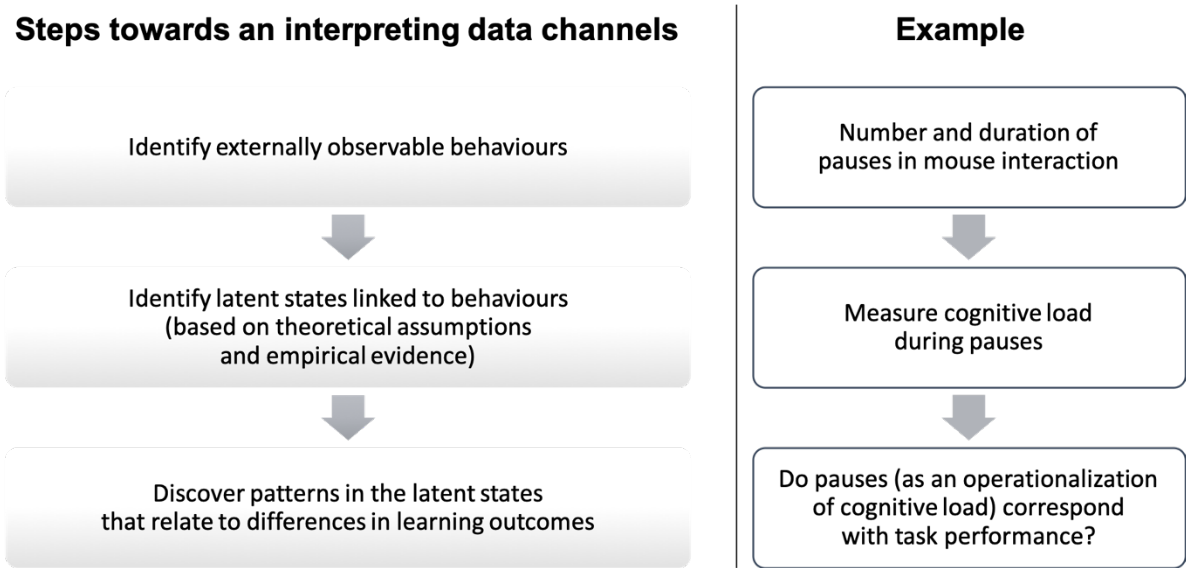


Figure 4. Steps towards interpreting data channels regarding latent variables.

Step 2: Identifying latent states linked to behaviors

In a second step, it needs to be checked whether the identified indices correspond with a latent construct. While it is a plausible and common practice to operationalize constructs as measures that intuitively appear as closely related indicators, this can still lead to false positive inference. As an example, time on task is often used as a measure for motivational persistence in literature (Vollmeyer & Rheinberg, 2000). However, time on task as measured in online learning environments could also indicate boredom if we are not aware of what exactly happens. Thus, triangulation of data channels might be needed, and deep theoretical knowledge as well as strong empirical evidence about the construct and the characteristics of the data sources are necessary to justify an operationalization.

Theoretical knowledge can give us hints on 1) which relevant constructs might be connected to the identified behaviors and 2) where those latent states might be in a vast array of information. As an example, in a study of this work (chapter 4.2) that investigates relations between writing and motivation, theories and empirical evidence on motivation serve to find possible latent states: it is known that there is a positive relationship between persistence and motivation (Vollmeyer & Rheinberg, 2000), and persistence is often operationalized as time being spent on a task (e.g., Nijstad, Stroebe & Lodewijkx, 1999). Thus, higher motivation could potentially be linked with behavior related to spending more time on the writing of text, such as typing longer texts (i.e., higher number of keystrokes) or more frequently revising a text (i.e., the number of corrections made on the text). This behavior can be quantified by the typing behavior. Compared to time on task, these detailed process measures represent significantly higher granularity.

Although theoretical assumptions are necessary, they are not sufficient to proof the validity of the operationalization. Additionally, concurrent data of established and valid measurements for the latent construct need to be linked with the behavioral indices in a reasonable way, following the logic of criterion validity (i.e., established test A measures construct B, so new indicator C measures B if C corresponds with A). In other domains, this step is often referred to as data labelling (Lali et al., 2014). This might be the most critical step towards interpreting data sources because invalid data labelling leads to the description of invalid links between the extracted behavior and wrong labels. However, one still needs to be aware that the solely proof of validity by correlating existing tests may lead to an invalid circular reasoning, if the existing measures is not reliable or valid. Besides this deductive method, linking data with existing measurements of latent states can very well be a way to inductively get new insights and build theories on them. For example, typing speed can be measured during a problem-solving process and code success rates. After that, both variables can be correlated to reveal a possible relationship between typing speed and problem-solving competence in a specific domain.

Step 3: Discover patterns in the latent states relating to differences in learning outcomes

In a last step, when behavioral indices as potential indicators for latent states were identified, it is then be checked whether patterns of those states relate to differences in learning outcomes in a third step, as existing theories would predict. For example, does the current motivation measured through typing behavior explain part of the variance of the learning performance?

2.5 Instructional Interventions in Technology Enhanced Learning

The studies of this work all aim at finally improving learning outcomes in advanced learning technologies, either indirectly through investigating data channels that can contribute to future adaptive learning systems, or directly through using data channels in an intervention. Therefore, the studies discuss the application of the results in the context of different instructional interventions, or experimentally explore different interventions itself. This chapter first introduces the idea of adaptive learning systems as a general framework for interventions, and then focusses on the actual interventions that have been used in the studies of this work: prompting, pedagogical agents, learning dashboards, and eye movement modeling examples.

2.5.1 Adaptive Learning Systems

The idea of adaptive learning systems is to support and enhance learning by fitting the presented environment to needs of learners, that are represented in different learner variables. Although in the studies of this work, no such system is investigated, the studies of this work aim at the idea of using collected data to inform adaptive intervention, i.e., peripheral data that accounts for the full interaction between the learner and the learning environment including the context information of environments and interfaces as described in chapter 4.1. Hence, it is important to understand the idea of adaptive learning systems and introduce a suitable theory for it. In their four-phase model, Shute and Zapata-Rivera (2008) present such a model of an adaptive cycle. It describes the process of adaptivity on the basis of the interaction of a learner with a digital system, and is depicted in Figure 5. The cycle consists of four components, namely: Capture, Analyze, Select and Present.

Firstly, the system captures data about the learner during the interaction in a learning environment. Examples of data that the system could record include mouse or typing behavior, eye movements, and physiological measures. This information forms the basis for the learner model that will be developed in the following. During the entire learning process, data is collected to update this model. The second step is to analyze the data obtained in order to create a first learner model based on the content-specific information of the learning environment. A suitable learner model ideally indicates the learner's current knowledge and the relevant deficits. This information is then used in a third step to decide for the need and the type of an intervention, e.g., a hint, an explanation, a specific behavior of an agent or a prompt. The selection of suitable interventions are the core of an adaptive system. Predefined decision rules and threshold values determine the suitability of the selection, which in turn can be dynamically updated as learning progresses. The final step deals with the presentation of the selected adaptive intervention measure. Although the described model initially has a linear course, regressions and regressive analyses are inevitable in the further course in order to keep up with the learner's developments. While the initial model may be rather coarse and unspecific, it ideally becomes more accurate over time. Thus, the learner model is not a static, but rather a self-updating, dynamic reflection of the learner.

2.5.1 Prompts

Prompts or prompting measures used in education and instruction are support mechanisms that aim to "induce or stimulate cognitive, metacognitive, motivational, volitional, and/or cooperative activities during learning" (Bannert, 2009, p. 140). In contrast to instructional content, prompts usually do not contain additional information, but support the application of already acquired knowledge or skills. Metacognitive prompts are a specific form of prompts aiming at activating metacognitive activities that are often needed in SRL (see Bannert, 2007, 2009 for an introduction).

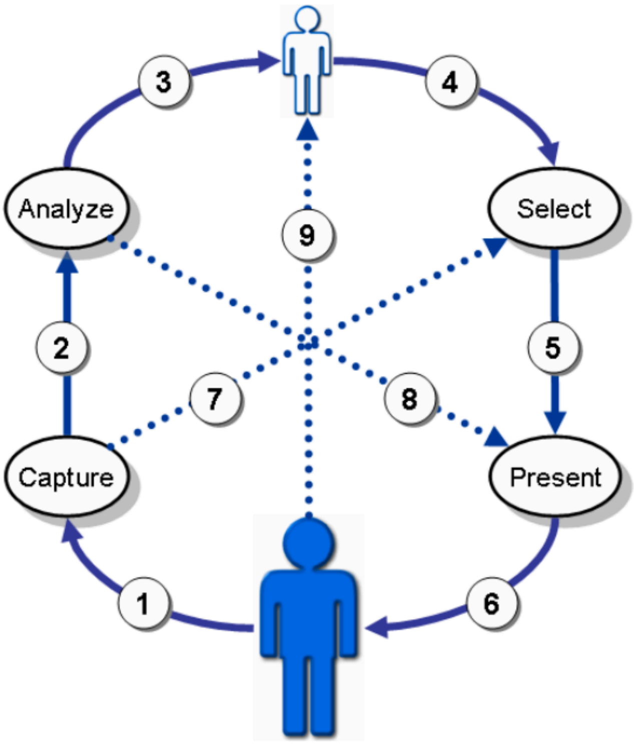


Figure 5. Cycle of an adaptive system as suggested by Shute and Zapata-Rivera, adapted from Shute and Zapata-Rivera (2008, p. 4)

The importance of such metacognitive activities and learning strategies is reflected both in theory and empirical investigations on SRL (e.g., Winne, 2001). Successful learners perform a range of such metacognitive activities. Even before the "actual learning", examples for metacognitive activities include analyzing the situation, orienting themselves by skimming task descriptions, or specifying learning (sub-)goals. While learning, learners need to judge the relevance of content for their goals, extract information and elaborate it. At the end of a learning activity, an evaluation of the achieved learning product considering their goals should take place (Bannert & Mengelkamp, 2013). Research in metacognition revealed that although learners have such metacognitive skills and know how to apply strategies, they often do not apply these spontaneously, leading to lower learning outcomes (e.g., Azevedo, 2009; Bannert, 2007; Bannert, Hildebrand & Mengelkamp, 2009; Zimmerman, 2008). This so-called "production deficit" (e.g., Bannert & Mengelkamp, 2013; Veenman et al., 2006) is the underlying assumption for metacognitive prompts. Thus, metacognitive prompts aim to trigger metacognitive activities by presenting learners with questions or statements asking learners at certain times during the learning process to reflect/monitor or control aspects of the learning content or their own mental activities. It is assumed that the resulting increased application of learners' repertoire of metacognitive activities will then enhance learning outcomes.

Prompts have a range of different parameters that need to be set, such as which learning activity should be prompted (e.g., Wichmann & Leutner, 2009), when should they occur (e.g., Thillmann, Künsting, Wirth & Leutner, 2009), which modality should be used to present them (e.g., auditory through an pedagogical agent, Azevedo et al., 2012), how specific should they be (e.g., Davis, 2003; Glogger, Holzäpfel, Schwonke, Nückles & Renkl, 2009), how should they be worded, or should learners be able to customize their own prompts (Bannert, Sonnenberg, Mengelkamp & Pieger, 2015; Pieger & Bannert, 2018). When using prompts in instructional aids, these parameters should be well chosen and based on empirical evidence. Although the effects of prompts are already well understood, there are still open questions regarding how learners interact with prompts (Bannert & Mengelkamp, 2013), which this work does in the last study (chapter 4.5). Answering these questions can potentially contribute to further specifying the optimal parameters for prompts in different conditions. Moreover, there are some general design principles for metacognitive aids that should be followed (Veenman et al., 2006) such as integrating metacognitive instruction into domain-specific instruction instead of teaching it without subject context, explaining why certain strategies are useful, and allow for sufficient training time in order to ensure that metacognitive activities can later be applied spontaneously.

2.5.2 Pedagogical Agents

Another suggestion to support SRL is the use of so-called virtual pedagogical agents. Although the effects of these systems on learning are not investigated in the studies of this work, agents were used in the learning environments. Therefore, it seems necessary to briefly explain the theoretical background and the empirical status of the systems. Pedagogical agents are mostly presented in human-like form within a virtual learning environment (e.g., Graesser, Wiemer-Hastings, Wiemer-Hastings & Kreuz, 1999; Johnson, Shaw & Ganeshan, 1998). The presentation varies from simple static images with visual text presentation to complex animated two- or even three-dimensional figures with speech input and/or output. The agents act as teachers, mentors, coaches, tutors or peers and provide cognitive, motivational and/or metacognitive support (e.g., Clarebout, Elen & Johnson, 2002). Although the idea of a virtual supporter has existed for decades, e-learning with "Human Computers as Co-Coaches" Erpenbeck & Sauter, 2013, p. 5), taking into account intelligent adaptivity through new technical possibilities, is also discussed in the current literature as an important form of teaching-learning and as a promising perspective (MMB-Trendmonitor, 2014).

Research in educational psychology has been investigates the effects of pedagogical agents since the nineties. Numerous studies focus on the effects of different appearances, forms of communication and response types of pedagogical agents, but not on the used instructional strategies (Dehn & Van Mulken, 2000). In most cases, perception and acceptance indicators were examined as dependent variables. Little is known about the effects on variables directly relevant to learning. In particular, a lack of studies confirming increased learning performance is described (e.g., Heidig & Clarebout, 2011). The benefit for learners is therefore controversial (e.g., Clarebout et al., 2002; Moreno, 2005). This is one of the reasons why critics complain that the high effort required to implement pedagogical agents is disproportionate to the benefits for the learner or that pedagogical agents even have a disruptive effect on the learning process (e.g., Chen et al., 2012; Choi & Clark, 2006; Clark & Choi, 2007).

Theoretically, the use of pedagogical agents is often justified by the creation of social effects in the learner. Based on the findings that interactions with computers can cause human social reactions (media-equation, e.g., Nass, Moon, Fogg, Reeves & Dryer, 1995), the persona effect was described in the context of pedagogical agents. According to the persona effect, the sole presence of a pedagogical agent promotes the learning process (e.g., Lester et al., 1997). Although the persona effect could not be replicated, it is still frequently quoted today (e.g., Craig, Gholson & Driscoll, 2002). The Social Agency Theory (also "Social-Cue Hypothesis") describes the assumption in pedagogical-psychological research that social cues from virtual agents lead to a pre-activation (priming) of social response behavior and consequently contribute to higher motivation and deeper cognitive processing of learning material (Mayer, 2005). The empirical results on the social-cue hypothesis are inconsistent. Although higher motivation and better transfer performance were empirically confirmed (Atkinson, 2002; Moreno, Mayer, Spires & Lester, 2001), it was later shown that these effects were probably due to the auditory text presentation of the educational agent in the sense of multimedia learning (Moreno, 2003). It was then postulated that only the voice of the educational agent was effective, regardless of its representation ("presence principle", Mayer, Dow & Mayer, 2003). Domagk (2008, p. 50) criticizes this, as she argues that pedagogical agents are defined through having a visual representation.

Regardless of the theoretical foundation, the questions of whether pedagogical agents promote motivation and learning and under what conditions they work were holistically investigated by Heidig and Clarebout (2011) in a meta-analysis of 75 articles. Of these, however, only 39 studies dealt with variables relevant to learning at all. Only 15 studies were designed as experiments with control groups without the use of an educational agent. The majority of the studies (9 out of 15) did not find any differences in learning success, motivation was only recorded in four studies at all, of which three showed no differences. Consistently reported were missing differences between animated, static and no agents with respect to recall performance (Baylor & Ryu, 2003; Dirkin, Mishra & Altermatt, 2005; Lusk & Atkinson, 2007). Only for the attractiveness of the agents consistent positive effects on the transfer performance could be shown in two studies (Domagk, 2010).

Most studies on the effectiveness of pedagogical agents compare different types of pedagogical agents without a control group (24 of 39 studies in Heidig & Clarebout's meta-analysis, 2011). Only an advantage of explanatory versus corrective feedback regarding transfer performance (Moreno, 2004; Moreno & Mayer, 2005) and an advantage of auditory versus visual text explanations (Atkinson, 2002; Craig et al., 2002; Mayer et al., 2003) are considered as proven here. After the publication of the meta-analysis, the so-called embodiment effect for the transfer performance could also be proven in three experiments. Based on the persona effect described above, this means that pedagogical agents with real gestures, facial expressions and language achieve better learning outcomes with learners (Mayer & DaPra, 2012).

2.5.3 Learning Dashboards

With the actions of learning analytics being described as the “measurement, collection, analysis and reporting of data about learners” (Gaševic, Dawson & Siemens, 2015, p. 1), learning dashboards emerged as a common intervention meant to enhance learning. Learning dashboards contain visual representations of data on learning processes that has been collected, processed and analyzed before. They are meant to “aggregate different indicators about learner(s), learning process(es) and/or learning context(s) into one or multiple visualizations” (Schwendimann et al., 2017, p. 8). Based on this definition, learning dashboards can be characterized by a range of questions or parameters: What data is represented to whom, when, and how?

What data is represented? – Every data stream that is regarded as relevant to the learning process can potentially be visualized in dashboards by researchers or instructional designers. However, as learning dashboards are mostly implemented in online learning environments, they often incorporate click-stream data (e.g., time spent on pages, number of logins into a learning management system) or performance data of tasks and quizzes. Data visualized in dashboards can be aggregated on a course level (e.g., mean score of the group for a task) or on an individual level (e.g., how an individual learner scored in a task). Moreover, it can be a comparison between the individual and the course level (e.g., compared to the course, a learner scored lower in this task) or between the individual and a standard (e.g., compared to a proposed or required standard, a learner scored higher in this task). The time range of the collected data can vary between a single learning session and a whole course term.

How is the data visualized? – Data visualization in dashboard aims at presenting students with a simple representation of sometimes complex data that is acquired during their learning. Examples are bar charts, pie charts or tables. Besides visuals, recent research is also tried to add automated explanatory texts to dashboards visuals (Ramos-Soto, Lama, Vazquez-Barreiros, Bugarin & Barro, 2015).

Who is the recipient of the dashboard? – While most learning dashboard are currently designed to be viewed by teachers or students (Bodily & Verbert, 2017; Schwendimann et al., 2017), other audiences can be administrators, study advisers or designers.

When is the dashboard shown? – Dashboards can be shown at different points in the learning process. Often, dashboards are presented as the first page after logging into a learning management system. In other systems, students need to explicitly click on the learning dashboard in a LMS, are presented with the learning dashboard at fixed time intervals during the learning process, or receive their dashboards as report e-mails.

For many of the current studies that examine or report about learning dashboards, it is criticized that these do not have a strong foundation on theories of educational psychology, as it is still a young research subject that emerged from the field of learning analytics (e.g., Gaševic et al., 2015). One consequence of this is that the actual pedagogical goal and the underlying psychological mechanism of the dashboard is often not clearly defined or not described at all. This can be argued for dashboards on a general level, but also for the specific type of data and visualizations that are chosen to be presented in the dashboard (i.e., what role does a certain presented information play for the mechanism of the dashboard?). When it comes to a specific type of visualization in dashboards, research in cognitive psychology on how different visualizations can impact the perception of information is often neglected in dashboard studies. This is an issue that has been recognized decades ago in other disciplines. As an example, in the domain of decision taking, Jarvenpaa (1989) already argued that "the designers of decision support systems lack theoretically based principles for designing graphical interfaces", and examined the effects of first computer-based graphical representation on information processing strategies. However, the decision of whether to use a bar chart to present information instead of a scatter plot should be based on empirical research, if available. Empirical research and theory development on the evaluation of visualization continues to be conducted in the fields of information visualization (Ware, 2013) and visual analytics (e.g., Keim et al., 2008; Nazemi, Burkhardt, Hoppe, Nazemi & Kohlhammer, 2015). Thus, this needs to be considered when designing dashboards. Such questions are a necessary part of the characterization of the dashboard if we look at it as an instructional intervention.

Another perspective is that researchers claim to "support awareness and reflection" through existing dashboards (e.g., 20 out of 26 studies in the review of Jivet, Scheffel, Drachsler & Specht, 2017). While this might implicitly mean that the dashboard supports metacognitive planning and monitoring by raising the awareness of the own learning process, authors often do not make this explicit and do not further specify what the pedagogical and psychological implication of this raised awareness should exactly be.

It is argued that the final goal of learning dashboards as instructional interventions should be a positive effect on the learning outcome. Explicitly defining the functions, goals, and mechanisms behind dashboards that lead to such positive effects is a prerequisite to get a rigorous picture of the impact that dashboards can have on learning. As an example, a learning dashboard of a vocabulary learning application could incorporate different functions with different goals and mechanisms: listing the items (function) that a learner should be able to recall (goal) is meant to support on a cognitive level (mechanism), or presenting statistics on how many items were recalled in the last session (function) should make users aware of their task conditions (goal) by supporting on a metacognitive level (mechanism). A slightly changed function can have a different goal and mechanism, e.g., showing a comparison between learners on how many items were recalled (function) could aim at supporting on a motivational level (goal) based on mechanisms described in theories of social comparison processes (e.g., Festinger, 1954).

Moreover, even considering studies on learning dashboards that are not strictly based on theories of educational sciences, there is no large body of empirical research yet. In the recent discourse and in meta analytic studies (Bodily & Verbert, 2017; Dawson, Jovanovic, Gašević & Pardo, 2017; Jivet et al., 2017; Schwendimann et al., 2017), exactly this lack of research on the actual effects that learning dashboards have on the learning outcomes is claimed. In Bodily & Verbert (2017), for example, only 2 of 94 papers examine actual effects on learning. It is claimed that we need further research on how learners interact with dashboards (e.g., Pardo, Poquet, Martinez-Maldonado & Dawson, 2017). It seems that the focus of current studies rather is on investigating student perceptions, technical functionality and different data sources – which is of course important to examine but is not sufficient for educational interventions. Existing studies that do measure effects on learning do not consistently report a positive impact of learning dashboards (e.g., Park & Jo, 2015) and therefore, estimating a more general effect size of learning dashboards is not yet feasible.

2.5.4 Eye-Movement Modeling Examples

Another recently developed way to support learning in multimedia learning are so-called Eye-Movement Modeling Examples (e.g., Gegenfurtner, Lehtinen, Jarodzka & Säljö, 2017; Jarodzka, van Gog, Dorr, Scheiter & Gerjets, 2013; Krebs, Schüler & Scheiter, 2018; Mason, Pluchino & Tornatora, 2015; van Marlen, van Wermeskerken, Jarodzka & van Gog, 2016). EMMEs reflect the eye movements recorded during learning in a technology-enhanced learning environment. These recordings are presented to learners as a model of how a particular task has been solved by others. EMME has been shown to help improve learning through better cognitive processing in multimedia environments (e.g., Scheiter, Schubert & Schüler, 2017). This improved cognitive processing is explained on the one hand by the idea that EMME activates a learner's prior knowledge of how information can be processed or that EMME leads to the acquisition of new processing strategies. Another rationale is that another person's eye movements represent a social cue that stimulates deeper cognitive processing (Krebs et al., 2018). Although the studies in this paper do not use EMMEs, the intervention study in this paper uses heat maps of mouse movements. EMME is relevant as a concept for these heat maps, because mouse movements and eye movements correlate (Guo & Agichtein, 2010; Huang & White, 2012; Huang, White & Dumais, 2011), have the same structure (X/Y coordinates over time) and eye tracking data is often presented in heat maps as well (Špakov & Miniotas, 2007).

3 Research Questions Overview

As the description of the theoretical background of this work demonstrates, there is still a lack of appropriate instruments to depict processes of SRL in technology-supported environments and to adequately capture their cognitive, metacognitive, motivational and affective facets and make them available in real time for analyses and adaptive interventions. In addition to the traditional quality criteria of objectivity, reliability and validity, requirements for such process measures include in real-time ("online") recording with high temporal resolution and low obtrusiveness and reactivity. Previously used data channels such as protocols on thinking aloud, screen recording, eye tracking, log files, video observations or physiological sensors meet these criteria to varying degrees. A data channel that has received little attention in research in educational psychology, but is non-obtrusive, non-reactive, objective and available online, is detailed data on observable interactions of learners in online learning environments. This data channel is introduced in this thesis as "peripheral data". It records both the content of learning environments as context, and related actions of learners triggered by mouse and keyboard, as well as the reactions of learning environments, such as structural or content changes. Although the above criteria for the use of the data are met, it is unclear whether this data can be interpreted reliably and validly with regard to relevant variables and behavior.

The aim of this dissertation is therefore to examine this data channel from the perspective of SRL and thus contribute to closing the existing research gap. For this purpose, three research questions are formulated, which are to be answered with one development project and four empirical studies.

In the development part of this thesis (chapter 4.1), the theoretical and methodological characteristics of peripheral data are investigated and the development of a software for the acquisition of the data is described. Hence, it addresses the following question:

1) Is peripheral data a suitable data stream to record and analyze the interactions of learners with learning environments?

On this methodological basis, the first three empirical studies (chapters 4.2, 4.3, and 4.4) will investigate how peripheral data address the following question:

2) (How) is peripheral data linked to cognitive, motivational, affective and metacognitive states of learners?

Finally, the last empirical study will address the impact that the visualization of learning dashboard in combination with metacognitive prompts has on the learning outcomes in online learning environments. Thus, the following question is investigated:

3) Can learners benefit from presenting them with visualizations of their acquired peripheral data in learning dashboards?

4 Research and Development: Analyzing and Supporting Self-Regulated Learning through Peripheral Data

In this chapter, the empirical studies and the software development of this work are described. First, the ScreenAlytics software frame is introduced as a methodological basis for the following empirical studies. This includes the comparison of existing methods to record observable behavior in online learning environments with a suggested data stream called peripheral data. In contrast to the later documented empirical studies, the description of the software initially has a methodological focus and contains descriptions of the technical details and software features, as well as a performance evaluation. In the discussion of this software, potential applications in psychological research are briefly described. After that, three empirical studies are described that investigate the relation of this recorded information to variables that are relevant for SRL. The last study addresses the question whether learners can benefit from presenting them with the recorded data by supporting their metacognitive activities. For each of the studies, a brief theoretical background that identifies research gaps and deriving questions is given, as well as information on the methodological implementation of the study, results and hypothesis testing. Results of the studies and their possible implications and limitations are then discussed.

4.1 Developing ScreenAlytics: Methodological Basis for Empirical Studies

Today, researchers of many disciplines make use of modern web-technologies to implement (experimental) environments for collecting data from human participants. Web technologies and available tools that support researchers in collecting data (e.g., de Leeuw, 2015; Reips & Neuhaus, 2002) are not only applicable for experiments delivered through the internet but also for in-lab browser-based data acquisition (Hilbig, 2015).

Moreover, in field research, online data acquisition is getting more attractive and relevant from a methodological point of view as it has advantages compared to classical in-lab experiments, such as larger samples with a higher heterogeneity that can be recruited more quickly (e.g., Birnbaum, 2004; Reips, 2000, 2002; Skitka & Sargis, 2006). Especially in psychology, these advantages are crucial with regards to the replication crises and an often very selective sample (Henrich, Heine & Norenzayan, 2010). Recent research has successfully shown that web-based experiments are able to reproduce the findings of a range of classical in-lab experiments in psychology that are often based on reaction times (e.g., Hilbig, 2015; Semmelmann & Weigelt, 2016). John and Samuel (2000) could show that findings of online self-report questionnaires are consistent with findings from traditional offline methods. In such studies, researchers are often not only interested in the final outcomes of questionnaires and tasks anymore, but need to conduct analyses on detailed interaction between participants and web-based (experimental) environments. This interaction is reflected in process data. Examples for the use and necessity of process data can be found in many disciplines such as human-computer interaction (Tang, Liu, Muller, Lin & Drews, 2006), survey methodology (Horwitz, 2013; Horwitz, Kreuter & Conrad, 2017), social psychology (Freeman, Pauker, Apfelbaum & Ambady, 2010) and educational psychology (Sonnenberg & Bannert, 2016). As an example, in educational psychology and the research on technology enhanced learning, studies need to investigate not only the outcomes of learning but especially the learning processes in order to gain insight into underlying mechanisms, promote the construction and verification of learning theories, and foster learning.

4.1.1 Comparing Methods of Recording

Recording interaction processes is currently realized either by screen recordings that generate videos of the interaction, by mouse/keyboard tracking software that logs mouse activity and keystrokes, or by using more or less detailed log-files of the web server. These methods have inherent disadvantages that are discussed in this chapter and that are solved by introducing the recording tool ScreenAlytics, and the underlying approach.

4.1.1.1 Screen Recordings

Screen recording is a method that digitally captures the output of computer screens as a pixel-based video file during the interaction between a user and a computer. When combined with audio narrations for educational purposes, the method is sometimes referred to as screencast or screen capture (e.g., Lloyd & Robertson, 2012; Razak & Ali, 2016; Veronikas & Maushak, 2005).

Regarding research and evaluation, recording the users' screens as videos is a commonly and successfully used way to examine user behavior in computer based environments (Tang et al., 2006). This is not surprising as recording the sessions can provide us with replays of the complete interaction between the participant and the corresponding context such as a web environment or a system application. Thus, it can reveal relevant information about user behavior. As an example, in our research on technology enhanced learning, researchers can observe and analyze how learners use instructional support in web environments, how often they correct solutions in exercises, or how much time they spend on viewing materials such as texts, videos or illustrations.

However, what can be acquired from widespread screen recording software like Cam Studio (http://camstudio.org) or VLC (http://www.videolan.org/vlc/) are pixel-based video files (e.g., mpeg, mov, avi). This output format is crucial as producing pixel-based videos results in data that is no further automatically processable or interpretable and therefore requires substantial work to be analyzed. As an example, although it can be seen what a user typed in a screen recording video, the text cannot be extracted automatically. Every single video needs to be watched manually to code the text (or use costly computer vision for text recognition needs to be implemented) in order to conduct further analyses on it. Another example is the analysis of the usage times and frequencies of specific elements or dialogs in web-based environments. It can be observed in the video but, to analyze, one must count and code it by hand.

Thus, by recording pixel-based videos, information is lost that actually has already been available. To retrieve that information, researchers have to conduct time-consuming manual coding. Besides, manual coding is prone to inducing subjectivity bias and errors, and therefore, cross-validations of more than one coder are often needed. Many examples in educational research report about the vast workload that results from manually coding screen recordings. As an example, Zhang & Quintana, 2012 reported "year-long processes of repeated viewing and transcribing of the videos," (p. 187) or Yew & Schmidt, 2012 analyzed "around 70 hours of screen recording as each student was online for about 7-8 hours." (p. 384). Figure 6 shows the workflow of traditional screen recording including the manual coding process and the possibilities for analyzing the resulting filtered event data.

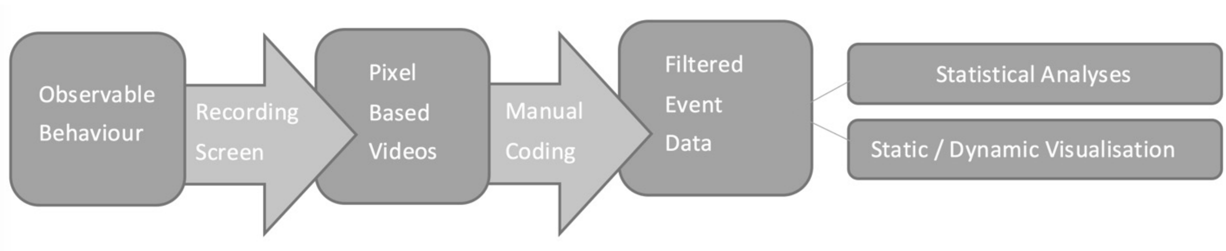


Figure 6. Workflow of Analyzing Traditional Screen Recordings

Aside from manual coding, other issues of screen recording advise against using it in web environments. Dependency on a client-sided installed software makes field-research outside the lab very difficult. Depending on the used software, screen recording can also have some obtrusiveness (e.g., a blinking red dot in the task bar) that may bias user behavior. Moreover, huge file sizes make it difficult to handle and archive resulting data and screen recording can be expensive regarding CPU and hard disk overhead, though having a negative impact on the computers' performance (Shea, Liu, Ngai & Cui, 2013). In addition, Tang, Liu, Muller, Lin &

Drews (2006) report strong privacy concerns of participants as screen recording is not limited to the interactions inside the application of interest (e.g., a web environment) but also captures the complete interaction in other programs (e.g., e-mails, private files, the desktop).

4.1.1.2 Mouse / Keyboard Tracking

Mouse and keyboard tracking systems record events triggered by the users' devices together with a timestamp. Installed either as a browser plugin or as a service on the operating system level, software packages capture mouse activity (i.e., x/y coordinates on the screen) and keyboard activity (key down and key up events) to log files or databases. This approach has existed for decades (e.g., "Input logger" by Trewin, 1998; "Tracer" by Lahl & Pietrowsky, 2008) and has been further developed into more sophisticated packages such as "Mousetracker" by Freeman & Ambady (2010) or "Mousetrap" by Kieslich & Henninger (2017). Thus, different disciplines successfully use mouse and keyboard tracking, for example in research on writing (e.g., "InputLog" by Leijten & Van Waes, 2013; Van Waes, Leijten & Van Weijen, 2009), usability research (Atterer, Wnuk & Schmidt, 2006) or social psychology (Freeman & Ambady, 2009; Freeman et al., 2010).

As the recorded data can be processed without manual coding, mouse and keyboard tracking solves a major issue of screen recordings by making the raw data accessible. However, these approaches still face important disadvantages. First, tracking applications still have to be installed as additional software on the users' computer, thus making it very difficult to use in field research. Although recording mouse and keyboard activity in web-based environments without additionally installed software is described in literature (e.g., Arroyo, Selker & Wei, 2006; Atterer et al., 2006; Mueller & Lockerd, 2001), the authors do not provide a tool to do so. Secondly, mouse and keyboard tracking software usually ignores the context in which the activity was recorded. Resulting data only contains raw event data (e.g., x/y coordinates of mouse movements together with a timestamp) without information about the context in which users showed the recorded behavior.

4.1.1.3 Log Files

Beside screen recordings and mouse/keyboard tracking, log files are commonly used in web-based research. Traditional log files are reports about requests to websites that are generated by and stored on a webserver, and were used for debugging since the very beginning of the internet (Suneetha & Krishnamoorthi, 2009; W3C World Wide Web Consortium, 1995). Although these (often cryptic) files can be analyzed with software tools like the LogAnalyzer (Reips & Stieger, 2004), it still provides only rough information about the interaction from which statements in the form “at time T, page P was visited by computer C” or more aggregated, “page P was visited N times” can be inferred. As server log files implicate other disadvantages (e.g., hurdles to identify unique sessions, see Zorrilla, Menasalvas, Marín, Mora & Segovia, 2005 for an introduction), more sophisticated web analytics software packages like Matomo (https://matomo.org, formerly known as “Piwik”) have been developed. These tools use additional client-sided information, often generated in JavaScript, to acquire log data that provides a more detailed insight into user behavior. The granularity can vary from recording the number of page views to detailed event information about users’ interactions. However, this requires researchers to customize the source code of their web environments in order to setup the recording of relevant events that they are interested in. Accordingly, usage cases of log data in psychological research range from simples descriptive analyses of navigation patterns to complex attempts of predicting users’ latent state variables like motivation from it (Cocea & Weibelzahl, 2007a, 2009).

Despite the limited information that log files can provide, recording it is straightforward: no software needs to be installed on the client computers and user-friendly tools like Matomo help researchers to acquire and visualize data. The resulting data can be exported and aggregated or directly used for quantitative analyses in statistics software. However, log data only provides static snapshots of an interaction at specific pre-configured events. This means that a researcher needs to configure each event (e.g., a click on a button of a website) that should be tracked – while ignoring what happens between the captured events and its contexts again. These drawbacks are crucial especially for explorative and qualitative studies that focus on generating theories or discovering meaningful patterns, because researchers need to have hypotheses in order to decide which events are tracked before the data acquisition.

4.1.2 Peripheral Data Combines Advantages of Other Measures

Considering the disadvantages of traditional screen recordings, mouse / keyboard tracking and traditional log data, an approach that solves these issues is proposed and introduced as *"peripheral data."*. Table 1 lists the advantages and disadvantages of the mentioned approaches to data acquisition in web-based environments.

Peripheral data is a processable documentation of the full interaction between a user and a web-based environment, including detailed information about a user's actions and the reactions of the environment to these actions.

Recorded user actions are mouse/touch clicks and movements, keystrokes, window scrolling, resizing, and (de-)focusing. The captured data is comparable to what we get from mouse and keyboard tracking software, but provides us with additional contextualization of that information. For instance, instead of just getting the x/y mouse position of a mouse click, one also knows on which element of a website the user clicked. Or, instead of just getting a keystroke, it is also known into which input element that key was typed. This contextualization is crucial as without it, no inferences regarding the content would be possible.

Table 1

Advantages and disadvantages of currently used approaches.

Approach	Advantages	Disadvantages
Screen recording	Records all available information, including context and behavior	Data is not further processable without manual coding; time-consuming; huge file sizes
Mouse / keyboard tracking	Processable raw data	No information about context and content, software dependencies on client computer
Log-files	No software dependencies on client computer	Low granularity, special events need to be configured

As every standard web browser allows us to observe these events, it is possible to record them via JavaScript event listeners (e.g., Alimadadi, Sequeira, Mesbah & Pattabiraman, 2014) without software dependencies on the users' computer. These event listeners return parameters like the x and y position when a user moves the mouse at a sampling rate of around 60 hertz or the width and height of the browser when it is resized. Together with a timestamp accurate to the nearest millisecond, these events can be sent to a server application asynchronously (i.e., without influencing the performance of the recorded environment), which stores them in a server-sided database. Hence, data structure is comparable to other approaches like traditional log-files or mouse tracking. Figure 7 shows the structure of example peripheral data events.

```
"move_move":{
    "id":145,
    "timestamp":1035,
    "x":12,
    "y":50,
    "target":"eq(120)"
},

"keystroke":{
    "id":157,
    "timestamp":2937,
    "charcode":79,
    "target":"#question"
},

"move_click":{
    "id":157,
    "timestamp":1537,
    "x":140,
    "y":100,
    "target":"#btn"
},

"scroll":{
    "id":157,
    "timestamp":4042,
    "x":703,
    "y":0,
    "target":"window"
}
```


Figure 7. Data Structure of Example Events from Peripheral Data

However, this data only reflects the actions of the users and not the reactions of the web environment to it. Hence, in addition to the mentioned input events, the approach also observes the initial web contents and changes on it over time, representing the reactions of the environment (using the DOM Mutation Observer that is implemented in all modern browser frameworks, see Mozilla Development Network, 2015). Again, this content can then be sent to a server application and stored on a server-sided database together with a timestamp accurate to the nearest millisecond.

As a result, the peripheral data approach allows both tracking the actions of the users and the reactions of the web environment, thereby representing the complete interaction process. This allows the later reconstruction of all actions and reactions so that researchers can view video-like replays of the complete interaction while still having access to the raw and processable data for further visualizations (e.g., heat maps) and quantitative investigations (e.g., analyses of detailed interaction data such as the typing behavior). The technical details of recording and replaying events are explained in the following sections of this chapter.

Recognizing and recording this data is both unobtrusive and therefor non-reactive. Regarding dependencies, no special client-sided hard- or software is required other than a standard web browser with JavaScript support (met by 99% of Web users in 2008; Kaczmirek, 2008, p.87) and a connection to the internet. Figure 8 illustrates the workflow of using peripheral data to record, visualize, and analyze interactions in web based environments.

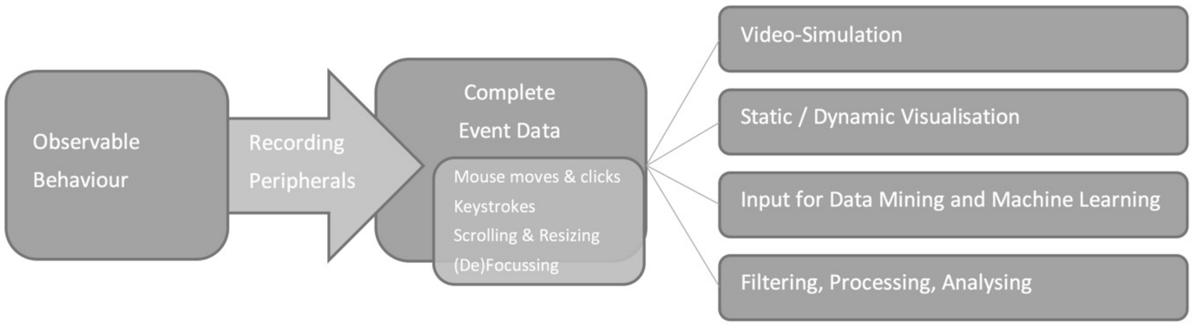


Figure 8. Workflow of Using Peripheral Data to Visualize and Analyze Web Processes

Although several commercial software packages exist that seem to implement the peripheral data approach (e.g., https://mouseflow.com or https://hotjar.com), the use of these for research purposes is very limited. Reasons for this are that the documentation of the underlying approach is not available to researchers and, most importantly, the software packages do not allow to access the raw data. This crucially limits the advantages of the tools as videos need again to be watched manually in order to extract relevant data such as content typed into text forms or interaction with a specific DOM element of interest. For recording mouse activity in online environments, the software “SMT” (Leiva & Hernando, 2007; Leiva & Vivó, 2013) is known. The major drawback of this software is that is does not account for changes in the DOM structure of the website, meaning it does not reflect reactions to the users actions or to user-specific content (e.g., when a user is logged in or assigned to an experimental / control group in an experiment). Moreover, the software seems to be no longer maintained and researchers need to have their own server infrastructure to use the software, hence requiring substantial technical skills or support to setup an experiment. Intensive search at the time of writing this work did not result in any software that implements the proposed approach.

4.1.3 Features of The ScreenAlytics Software Framework

ScreenAlytics is a software framework developed in this work that aims at supporting researchers with recording, visualizing, and analyzing web-based process data. It currently involves the following features: 1) recording and storing user actions and website reactions, 2) video-like replays of the web sessions including activity charts, 3) heat maps of mouse movement and clicks, 4) visualization of navigation paths, 5) extraction of text input and analyses of typing behavior as well as 6) custom event labelling and 7) an API. Features are described and reasons are given why they are helpful for researchers. ScreenAlytics is delivered to researchers as software-as-a-service. This means that ScreenAlytics runs on a remote server provided by the Technical University of Munich so that researchers do not need to have special technical skills or support in order to use the software. Prior to using the software, an agreement is submitted that all participants must be made aware of the data that researchers are collecting through ScreenAlytics and that no one besides the researchers will access the collected data.

4.1.3.1 Recording Interactions

In order to collect the described peripheral process data, ScreenAlytics can be embedded into any web-based environment. Therefore, after registering at the public ScreenAlytics online platform, a short JavaScript snippet is provided, which needs to be placed into the source code of templates or pages that should be recorded by the researcher. Figure 9 shows an example of a provided JavaScript snippet.

```
<script type="text/javascript">
    window._saq = window._saq || [];
    (function() {
        var sa = document.createElement("script"); sa.src = "https://tueds25-exp.srv.mwn.de/1.js";
        sa.async = true; sa.type = "text/javascript";
        document.getElementsByTagName("head")[0].appendChild(sa);
    })();
  </script>
```

Figure 9. JavaScript Snippet Provided by ScreenAlytics

When accessing a website that has been configured to be recorded, the ScreenAlytics client-sided tracking system is loaded from external servers and initializes JavaScript listeners for all mouse, keyboard, and window-related events as well as for changes on the *Document Object Model* (DOM, see W3C World Wide Web Consortium, 2005) of the website which reflects the content of the website. Those tracked events are then serialized, compressed, encrypted, and sent to a server via a secured Websocket or AJAX request every second (see Mozilla Development Network, 2016). The backend server application stores the information on a server-sided database. Both, the backend application and the database, is provided by the ScreenAlytics server, which is based in a data center of the Technical University of Munich. Figure 10 describes the process of tracking and storing JavaScript events and lists the tracked events.

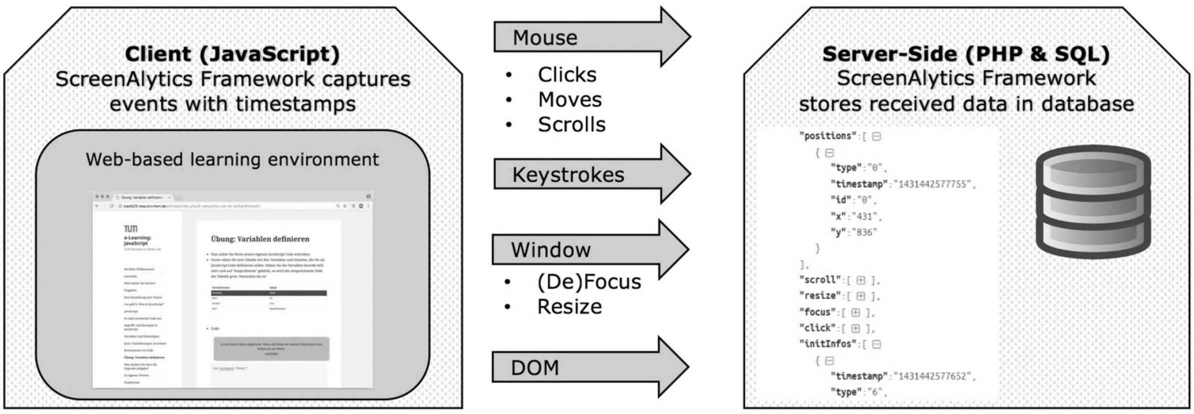


Figure 10. ScreenAlytics Captures Client-side JavaScript Events and Sends Them to the Server-side Database.

4.1.3.2 Video-like Replays of Recorded Sessions

ScreenAlytics provides researchers with different viewing applications. Firstly, recorded sessions can be replayed just like video based screen recordings. To achieve that, the viewer application reconstructs the initial DOM using archived versions of the website assets (e.g., images or stylesheets) of the recorded session. Then, it simulates all captured events including typed texts, navigation, clicks, scrolling, and resizing of the browser. The website is loaded within an iframe element, and mouse traces are displayed either as a continuous scan path (as often used for eye tracking, Harper, 2009) or as a simulated moving mouse cursor in an overlay. Researchers have access to a control panel, where projects, sessions, and specific pages can be selected. Using a slider, researchers can jump to specific timestamps within a recording and change the speed of the replay. The website is displayed in the same size as the visitor experienced it, but researchers can zoom in and out, e.g., when watching recordings of mobile devices with small display sizes. An event charts in the control panel represents frequencies of mouse, keyboard, and navigation events over time. Figure 11 shows the control panel and Figure 12 an example visualization of mouse moving in a learning environment.

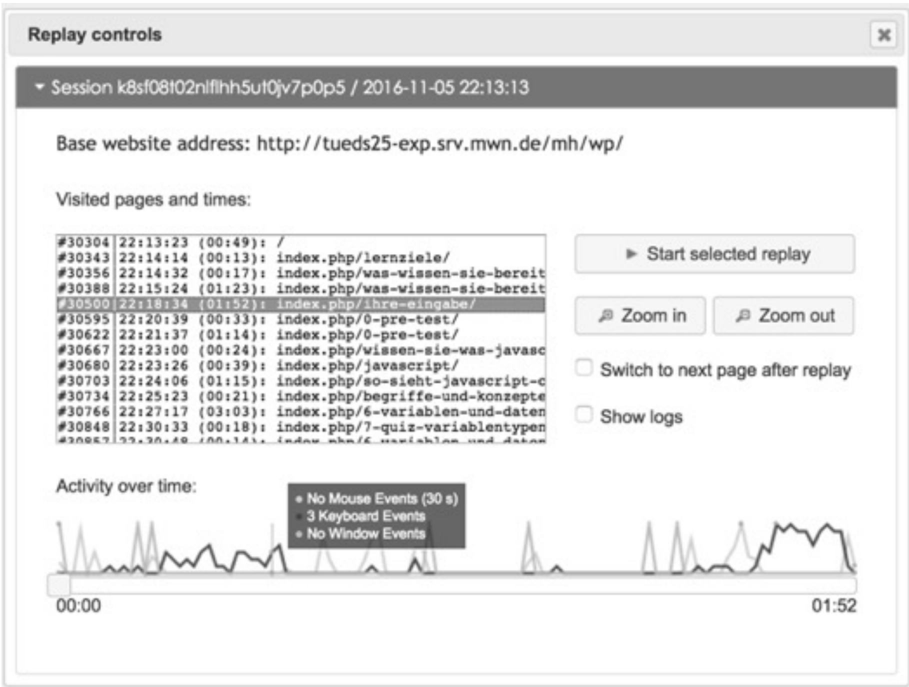


Figure 11. Researchers can select pages accessed by a user and view the activity of it in the control panel

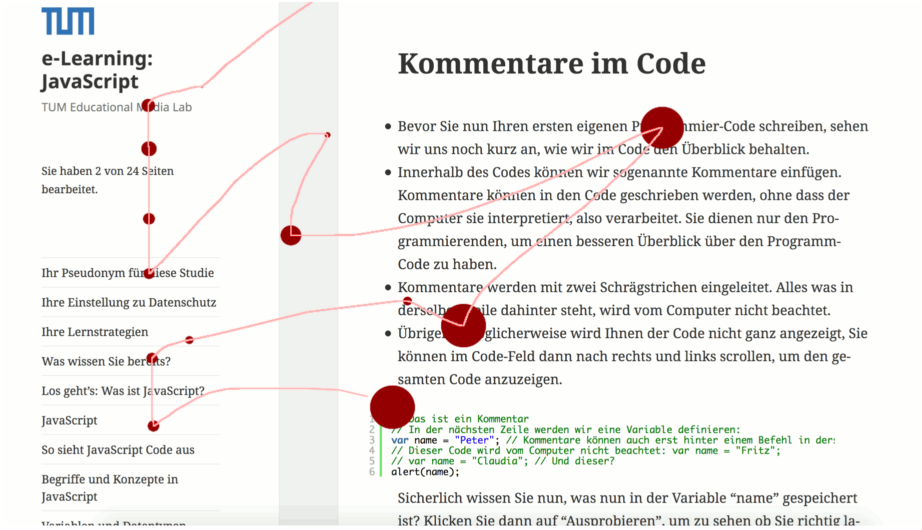


Figure 12. Visualization of Mouse Movements from Peripheral Data in a Learning Environment About Web Programming

4.1.3.3 Heat maps

Heat maps visualize the frequency x/y coordinates on computer screens graphically by using a spectral color continuum from usually green (minimum) via yellow (medium) to red (maximum). Heat maps are often used as a visualization in eye tracking research and indicate fixation counts or fixation duration either for a single person or for an aggregated group (i.e., to which extent have areas been focused by a person/group; see Špakov & Miniotas, 2007 for an introduction). This concept can be used for mouse behavior in the very same way, as there is no difference in data structure. Although it should be noted that the usage of heat maps (for eye tracking and mouse tracking data) to infer valid conclusions is a contentious topic (Bojko, 2009), three different heat map types to visualize mouse movements and clicks as well as for scrolling were implemented.

Regarding mouse movements, research has shown that there is a moderate to strong positive correlation between mouse cursor position and gaze position in general (Chen, Anderson & Sohn, 2001; Cooke, 2006; Guo & Agichtein, 2010; Huang & White, 2012; Huang et al., 2011), and that the correlation is higher during active mouse movement (Hauger, Paramythis & Weibelzahl, 2011). Moreover, click heat map was implemented as Huang and White (2012) found the smallest gaze-cursor distance when clicking on elements or links (a median of 74px). The click heat map also allows to see how many clicks have been registered on a specific HTML element.

When the page height is bigger than the browser height, users need to scroll in order to see an element which is below the fold. Heat maps were implemented that represent scrolling behavior in order to provide researchers with information about what percentage of a user group saw content that needs scrolling. Scrolling heat maps also help to check if a specific user saw an element (outside the fold) without watching the whole recorded session. These inferences are valid as scrolling is necessary to see elements below the fold.

In ScreenAlytics, researchers can create heat maps filtered by the type of interaction (move/click/scroll), specific sessions, and pages. Figure 13 shows a heat map of the aggregated mouse movements of an user sample working on a performance task in an online learning environment.

Figure 13. Aggregated Mouse Movements shown in a Heat map

4.1.3.4 Visualization of Navigation Patterns

Users' navigation behavior in web based environments has been utilized for research in broad range of disciplines, for example in the field of technology enhanced learning (e.g., Bannert et al., 2015; Graf & Liu, 2010; Puntambekar, Sullivan & Hübscher, 2013). Hence, ScreenAlytics provides a visualization of navigation patterns including the accessed page, the time spent on the page as well as information about the direction of the navigation (back to already visited page vs. first visit). Figure 14 shows an example of navigation patterns of three users. The tool also enables researchers to easily filter the visualization by pages and sessions and to export the navigation data for further analyses.

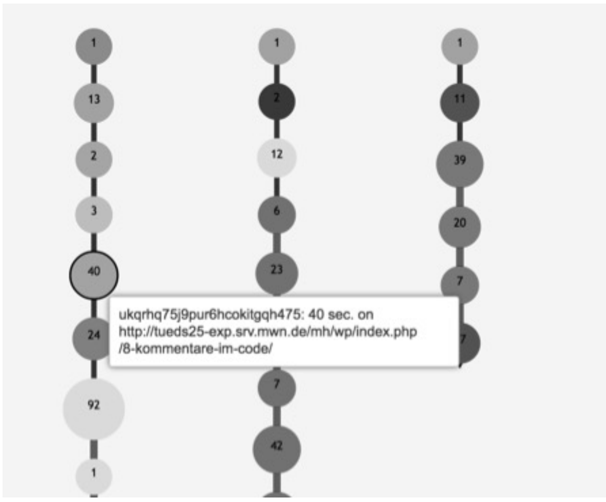


Figure 14. Visualization of navigation patterns of three web sessions. Colors indicate the website, numbers, and radius of the circles indicate the seconds on a page. Backward movements are indicated as colored connections.

4.1.3.5 Text Analyses

Regarding the analysis of text inputs, ScreenAlytics provides an automated recognition of all text input fields that are available on the recorded pages. Researchers can then select sessions, pages, and the input fields they want to analyze. ScreenAlytics creates an overview of all input activities as well as a simulation of the typing process in real-time. Figure 15 and Figure 16 illustrate examples of using the text analysis tool for investigating what learners typed in a learning environment about website programming. The researcher can see what text has been typed into the field and get information about indices of typing behavior (e.g., number of deletions, duration of pauses, average typing speed). The typing process can also be replayed as a video-like simulation. Preprocessing of data about typing behavior was implemented as

1) in research on technology-based learning, this is helpful for investigating how an answer to a quiz or task has been developed by the learner and 2) typing behavior is crucial to a range of other disciplines such as research on writing where excellent systems are only available for offline use (Leijten & Van Waes, 2013), research on authentication (Bergadano, Gunetti & Picardi, 2003), or research on the recognition of certain psychological latent variables (e.g., Epp, Lippold & Mandryk, 2011; Leong, 2016).

Figure 15. Text analyses are supported by automated recognition of text input fields on recorded pages.

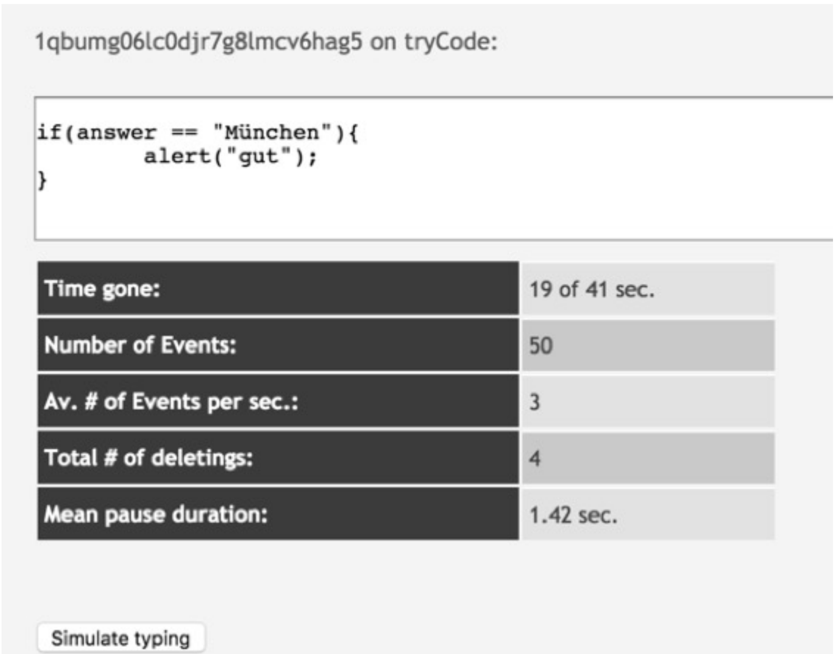


Figure 16. Simulation and Statistics of the Typing Process

4.1.3.6 Custom Event Labels

During the video-like replays, researchers can attach custom text labels to a timestamp of a session. This can be used for expert coding of theoretically important events, for example, in research on technology enhance learning, it can be used to investigate the use of metacognitive strategies during learning. Researchers can export the labels for analyses in statistical software packages.

4.1.3.7 API Functions

An application programming interface (API) has been implemented in order to 1) export data to other applications such as tools for data analysis and 2) store and read custom variables. As an example, for custom variable tracking, in intervention studies in technology enhanced learning, current achievements of learners in different tasks and quizzes can be stored as custom variables through the ScreenAlytics API in order to present learners with an overview of their learning processes at a later point. Hence, building an additional database and script to save this information is not needed. Moreover, researchers can easily export these states for further analysis. Figure 17 shows a sample API request from the statistics software *R*.

```
domain = "tueds25-exp.srv.mwn.de"
token = "556646343*****"
secret <- jsonlite::base64_enc(paste(domain, token, sep = ":"))
req <- httr::GET("https://ueds25-exp.srv.mwn.de/api/{userId}/{websiteId}/",
                  httr::add_headers(
                    "Authorization" = paste("Basic", gsub("\n", "", secret)),
                    "Content-Type" = "application/x-www-form-urlencoded;charset=UTF-8"
                  ),
                  body = "variable1=value1&variable2=value2"
)
json <- httr::content(req, as = "text")

```

Figure 17. Sending an API Request to ScreenAlytics from R.

4.1.4 Technical Evaluation

It was checked whether the implementation of ScreenAlytics affects the performance of websites. As ScreenAlytics requires the implementation of an external JavaScript library, slightly longer loading times of websites using it are expected. Hence, the tool WebpageTest.org (see Viscomi, Davies & Duran, 2015 for an introduction) was used to measure the effect that ScreenAlytics has on the loading of a standard Wordpress (https://wordpress.org) based learning environment used in an experiment in educational psychology. WebpageTest.org provides several metrics (described in more detail in WebPageTest.org, 2018) for testing the loading process. Firstly, “Load Time” was used, which is “the time from the start of the initial navigation until the beginning of the window load event.”, and where the “window load event” is triggered when the requested page as well as all externally resources are loaded. Secondly, the index “StartRender” was compared, which is “the time from the start of the initial navigation until the first non-white content is painted to the browser display.” (WebPageTest.org, 2018) “Load Time” was expected to be affected as external resources are loaded, but “StartRender” value was not expected to be increased significantly as ScreenAlytics is requested asynchronously (i.e., not blocking the rendering of a website). A total of $N = 52$ loadings were measured for both conditions ($N = 26$ with and N=26 without ScreenAlytics). The configuration of WebPageTest.org was set to *Connection = DSL (1.5Mbps 50ms RTT), Test Location = Frankfurt, Germany – EC2* and *Browser = Chrome*. As expected, “Load Time” for the condition with ScreenAlytics (M = 3516, SD = 129 [ms]) was significantly higher than without (M = 2764, SD = 254 [ms], T (52) = -.13,670, $p < .01$). However, “StartRender” was not significantly increased for ScreenAlytics (M = 2014, SD = 523 [ms]) compared to the control condition (M = 1959, SD = 415 [ms], T(52) = -0,432, $p = .667$). This means that a slightly higher loading time of M_{diff} = 752 ms will only affect users if the website depends on the “window load event”.

4.1.5 Usage Scenarios

Using peripheral data holds a great potential for exploring and generating data as well as validating theories. By visualizing the data as video simulations, relations and patterns can be detected more easily than in the complex structures of “raw” data (see Bowker et al., 2013 for an extensive discussion on the term). However, once such patterns are assumed, hypotheses can immediately be checked by traditional statistical analyses or machine learning algorithms on the same dataset with the available objective and non-biased raw data. The proposed software can be used for a wide range of both interventions and research. Besides the obvious application of usability testing of websites, there are promising usage cases that are further described: checking the data quality in online experiments, the video-cued recall method, using recorded data to foster learning, and modelling latent psychological variables from peripheral data.

4.1.5.1 Data Quality in Internet Experiments

Online experiments that are distributed via mailing lists or social networks are able to quickly recruit a large sample with a high heterogeneity. However, even advocates of web based research methodology claim that “this mode of research has some inherent limitations due to lack of control and observation of conditions” (Reips & Birnbaum, 2011, p. 563). It is argued that the data recorded by ScreenAlytics can help to reduce these limitations. Although a systematic usability study is needed to demonstrate the effectiveness of it, the following checklist was already used as a strategy to check the quality of participations in previous experiments.

1. Does the duration of the session deviate extremely from the mean duration?
2. Are there focus/blur events that indicate that a participant left to another window/tab and returned to the experimental environment?

3. Does the activity chart of the session indicate salient pauses while taking part in the experiment?
4. Is the device, the screen resolution and browser size that the user accessed the experiment with incompatible with the environment?

If one of the mentioned points were answered with yes, the actual recording was watched to decide whether or not that participant needs to be removed from further analyses.

4.1.5.2 Video-cued Recall

Another research application could be the use of ScreenAlytics for the video-cued recall method (e.g., Miller, 2004). Video-cued recalls aim at reducing the bias of self-reports by encouraging participants to view videos of their behavior. Thus, after finishing an experiment, participants could be asked to report about behavior that a researcher is interested in while watching (parts of) the recordings of his/her session as a cue (e.g., learners are asked to report on their usage of self-regulation strategies in their learning processes). This has already been suggested for eye tracking data (e.g., van Gog & Scheiter, 2010). Unlike traditional screen recordings, intelligent filters could be applied to select specific scenes of interest. For example, one could only select and replay scenes in which users navigate to a specific page, in which they typewrite, or in which they pause their interactions.

4.1.5.3 Using Peripheral Data to Foster Learning

There are several ideas on how learning in technology enhanced environments can be supported through the approach that ScreenAlytics uses. Firstly, peripheral data can be used to generate simulated scaffolds for learners that represents learning behavior or problem-solving processes. As mouse movements, clicks, typewriting, etc. can be simulated during the learning process, it would be possible to equip pedagogical agents (i.e., virtual characters that are designed to support learning processes, e.g., Veletsianos & Russell, 2014) with the ability to actively engage with the learner's screen. Simulated scaffolds could either come from previously recorded behavior of a didactic domain expert or could represent simulated worked-out examples. Another approach could be the presentation of complete recorded learning sessions of experts to enhance SRL. This follows the rationales of EMMEs which was presented in chapter 2.5.4.

The rapidly growing field of learning dashboards is another potential application (e.g., according to the framework proposed in Verbert, Duval, Klerkx, Govaerts & Santos, 2013). Learners could be provided with information on how their own interaction with learning environments differ from other learning sessions or specifically successful learners and give adaptive recommendations (e.g., "You spent only 2 minutes on page XY while successful learners normally work about 10 minutes on that page – do you want to review that page?").

4.1.5.4 Peripheral Data as Proxy Measures for Latent Psychological Variables

Besides the discussed possible usage cases, peripheral devices are an unobtrusive and non-reactive data source that is potentially related to latent psychological variables and can be used as a proxy measure for these. Using mouse and keyboard data to model a variety of user information is not a new idea. For example, typing behavior and mouse movement (so-called *keystroke* and *mouse dynamics*) are commonly used in the field of identification and authentication (e.g., Bergadano et al., 2003; Jorgensen & Yu, 2011), mouse-tracking is popular in usability research (Atterer et al., 2006), and social psychologists successfully made use of the mouse behavior in order to assess subjects' tendency towards stereotyping (Freeman & Ambady, 2009; Freeman et al., 2010). Moreover, eye tracking experiments discovered a medium correlation (r = .58) between mouse and gaze position (Chen et al., 2001).

Although not a new idea, ScreenAlytics makes the collection of the data in online environments more convenient and standardized and, as mentioned before, the contextualized data can add parameters to the feature space that are not available in isolated mouse or keyboard data. In addition to that, there is still a large research gap in modeling variables in the field of technology enhanced learning. Hardly any research has been done that connects peripheral data with learning outcomes. Existing psychological theories can give us hints on 1) **which** relevant latent states might be hidden in the identified behaviors and 2) **where** those latent states might be in that vast array of information. Besides this deductive method, linking data with existing measurements of latent states can also be a way to inductively get new insights and build theories on them (e.g., McQuiggan et al., 2008).

4.1.6 Conclusions and Next Steps

Using peripheral data for recording and visualizing sessions in web-based environments has many advantages over traditional screen recordings and log data recording. Accessibility and processability of the behavioral data is not lost, and statistical analyses can be conducted easily without manually coding events. Besides video-simulation, peripheral data allows multiple ways of visualizing the data (e.g., heat maps, navigation trees) and bringing it to other software for further analyses (through data export or API). No specialized hardware or software needs to be installed, and the server-sided storage of the data facilitates the acquisition processes without the need to get the video files from client computers. Thus, it enables researchers to conduct and implement complex research designs with sophisticated methodology outside the lab. Compared to traditional screen recordings, peripheral data only needs a fraction of the storage space and is thereby interpretable at runtime. Due to the data structure, which is very similar to the structure of eye tracking or physiological data, synchronizing it with other data channels or labelling it for machine learning algorithms can be done more easily.

When the ScreenAlytics software has been presented at conferences in the area of educational psychology and online research methodology two years ago, feedback to the system and requests to use it were overwhelmingly positive. Researchers suggested possible usage cases in their fields and requested features of which some are already implemented while others are the development agenda. As one major disadvantage is its limitation to enclosed web environments where researchers can place the JavaScript snippet on, a browser plugin that allows the recording of any website is the next feature to be built. Moreover, implementation of enhanced possibilities to automatically extract interactions with DOM elements is planned as this specific task currently requires skills in data mining.

4.2 Study 1: How Typing Behavior Corresponds with Learning Outcomes and Motivation

Writing tasks are commonly used in technology enhanced learning environments by both researchers and instructional designers. Examples range from short open answers in domain-specific exercises (Yang, Zhang & Yu, 2017), to learning journals and protocols (Cheng, 2017; Nückles et al., 2009), and complex essays in second language learning (Godwin-Jones, 2018). Moreover, in the domain of computer science, tasks that require learners to write programming code are widely used in online courses (e.g., Király, Nehéz & Hornyák, 2017). With the continuous growth of massive open online courses (MOOCs, see Shah, 2015) having thousands of learners enrolled in a course, there is a urgent demand for automated analysis of learners' texts in order to provide meaningful cognitive feedback or grade submissions. Researchers and instructional designers continue to struggle with such automated feedback or adaptive systems on written input of learners, because analyzing this input currently requires complex, content-depending, labor-intensive and inflexible algorithms that extract the meaning of texts, and interpret it regarding a very specific task. Although there has been huge progress in using artificial intelligence and machine learning for text processing in the last years, big training data sets are needed to achieve acceptable results that are still very specific regarding the task, content and domain. As an example, in conversational agent systems, simple algorithms are often implement as rule-based recognition of specific cueing words with randomly chosen answers of a previously defined set, which is why such systems are still in the “uncanny valley” of not being accepted by learners as an adequate conversational partner (Schönbrodt & Asendorpf, 2011; Shiban et al., 2015). Another branch of research focuses on the prediction of demographic or latent psychological variables through text mining. As an example, Kucukyilmaz (2006) used text-mining algorithms on chat messages to predict the gender of users, reaching prediction accuracies up to 84%. Another example is done by Anjewierden, Kollöffel, and Hulshof (2007), who tried to discover regulatory activities in collaborative learning environments using chat messages.

As it is very challenging to gain information for adaptive systems or the measurement of latent psychological variables from text content, another idea is the use of meta information about text and writing processes as an additional measure to improve prediction accuracy for proxy measurements of latent psychological variables. This can still be content-related meta data such, e.g., using the grammatical structures of texts to provide feedback on text coherence (e.g., Lachner, Burkhart & Nückles, 2017). However, meta data on writing processes can also be data on typing behavior, also referred to as keystroke logging or keystroke dynamics. Data on typing behavior typically describes events for pressing and releasing keys, including information on the type of key (e.g., characters, numbers, special keys such as control or delete) and a timestamp (see chapter 4.1.1.2). Typing behavior has already been used in other domains, e.g., identification and authentication (Bergadano et al., 2003). In research related to learning and instruction, first attempts are made to use keystrokes to distinguish frustrated from non-frustrated learners (Leong, 2016), detecting stressed learners in learning management systems (Lim, Ayesh & Stacey, 2014; Rodrigues, Gonçalves, Carneiro, Novais & Fdez-Riverola, 2013), detecting emotional states (see

Kolakowska, 2013 for a review) such as anger and excitement (Epp et al., 2011), or engagement and boredom (Allen et al., 2016).

However, when documenting this study, there was no research yet on whether typing behavior corresponds with learning outcomes and motivation. Thus, the idea of this study is to investigate whether meta-data of writing processes in programming environments can contribute to this discourse.

Typing behavior as a proxy measure for motivation

As motivation is an important factor in different models of SRL (e.g., Winne & Hadwin, 1998; Zimmerman & Moylan, 2009), it would be a major achievement to model the current motivation of learners through typing behavior. Motivation is broadly defined as an internal state that triggers behavior, controls the direction of it, and maintains it (for a detailed discussion on terms regarding motivation, see Murphy & Alexander, 2000). Being such a central prerequisite for learning, measuring motivation is important not only for further research in learning and instruction, but also to support and enhance learning. Especially for adaptive learning environments that react to learners' variables, a reliable and valid measure of motivation that is available in real-time is key.

This study focuses on the cognitive-motivational process model (Vollmeyer & Rheinberg, 1998) and the effects on SRL (Rheinberg, Vollmeyer & Rollett, 2000). The model describes current motivation depending on stable characteristics of the person (e.g., motives or interests) as well as on flexible characteristics of the situation (e.g., task difficulty or learning environment) that is related to the current or upcoming task). Rheinberg, Vollmeyer and Burns (2001) describe the four dimensions anxiety, probability of success, interest, and challenge to be specifically relevant for current motivation in learning situations. In their model, they describe time on task and quality of performed learning activities as variables influenced by the current motivation and as mediators of the learning process.

At the moment, motivation is either measured as a direct self-report through questionnaires, think-aloud protocols, or physiological responses. Researchers also use indirect measures such as observable cognitive (e.g., recall) or behavioral (e.g., performance) responses that need additional measures in order to be interpreted regarding motivation (Touré-Tillery & Fishbach, 2014). Thus, current measures of motivation are hardly able to provide real-time measures without being reactive or disturbing the learning process. Thus, in this study, typing behavior is examined as a potential real-time, unobtrusive measure for motivation.

Previous research rather focused on detecting specific events that are related to motivation than on predicting the actual level of motivation that learners experience in terms of the introduced model and its operationalization. Cocea and Weibelzahl (2006, 2007b) used log-file analyses to detect disengagement by applying several data mining techniques. They reached accuracy rates of up to 87% in predicting disengagement of learners. However, their data labelling of learners being engaged or disengaged was based on subjective expert ratings of log-files (e.g., reading a page for less than 30 seconds is disengaged). Thus, their algorithms decided on the same criteria as the experts and there has not been any validation that those criteria are really related to the level of engagement. One could criticize that they did not predict engagement but only recognized iterative patterns in log-files. This stresses the importance of valid data labelling as described before. Alike, Vicente and Pain (2002a, 2002b) let participants view recorded interactions in intelligent tutoring systems and instructed them to fill detailed coding form regarding learner's motivational traits and states. Although not rated by the learners themselves, the data labelling was well-grounded on a motivational framework. McQuiggan et al. (2008) labelled physiological data (heart rate and electrodermal activity) with self-reports of self-efficacy and reached classification rates of up to 86.9%. As these previous attempts show that there is information about the described events and variables included in the observable learning process, and as keystrokes depict parts of this observable learning process, keystrokes seem to be a potential proxy measure for motivation that is worth investigating.

4.2.1 Research Question and Hypotheses

The rationale of recording and analyzing typing behavior is that the fluency and flow of writing is a proxy for underlying cognitive processes. This is why the focus of analyses is on different indices such as length or frequency of pausing, corrections, etc. (e.g., Leijten & Van Waes, 2013). Similar to speech, indices like the length of a pause are interpreted as measures of cognitive effort. As an example, studies have shown that the duration of pauses increases with the level of text units, i.e., pauses between words are shorter than pauses between sentences, whereas pauses between sentences are shorter than pauses between paragraphs (Spelman Miller, 2000; Wengelin, 2006). Moreover, corrections may relate to a discrepancy between the intention of the writer and his or her produced text so far (Leijten, Van Waes & Ransdell, 2010), but also to grammatical errors. Thus, as previous research indicates that pausing and revisions indicate hurdles during the writing process, it is argued that indicators of higher typing speed (i.e., less pausing, less revisions) might indicate less struggling in writing.

Thus, it is hypothesized that indices of higher typing speed while typewriting continuous text in the recall task is associated with higher recall performance, higher declarative, and higher procedural knowledge (*Hypothesis 1: Fast-Typing-High-Performance-Hypothesis*). Likewise, it is hypothesized for typing during coding exercises that indices of higher typing speed correspond with higher declarative and procedural knowledge (*Hypothesis 2: Fast-Coding-High-Performance-Hypothesis*).

Regarding motivation, explorative analyses of the correspondence with typing behavior are conducted, but no explicit hypotheses are formulated. However, as time on task and quality of performed learning activities are named as potential indicators for motivation (Rheinberg et al., 2001), it is argued that keystrokes represent more fine-grained indices of time on task, and that they might be a proxy measures of the quality of learning activities (i.e., texts) and thus, correspond with motivation.

4.2.2 Method

4.2.2.1 Sample and Design

In a correlation study, 43 undergraduate students (10 males; $M_{age} = 19.66$; $SD = 1.03$) majoring in media communication at a German university participated. All were enrolled in one of two parallel courses ($N_1 = 22$; $N_2 = 21$) dealing with the conception and development of digital learning environments. Participants received no incentives but learning contents (website programming) were part of their course curriculum.

The laboratory was equipped with 22 iMac desktop computers (21,5 inch display with a resolution of 1920x1080, tethered apple mouse and keyboard). Firefox was used as the web browser, and the learning environment was presented in full screen mode.

4.2.2.2 Learning Materials

Students had to learn the basic concepts, terms, syntax and properties of "cascading style sheets" (CSS), a common standard to style websites. Learning material was structured linear and consisted of 20 content pages including about 2200 words, two tables, two illustrations, 13 code examples and five interactive coding exercises. In interactive coding exercises, learners had to write CSS code to solve a given task (e.g., "Set the width of the image to 200 pixels"). The results of their code were presented below the text area when clicking on a "Try it!" button as well as verbal feedback was provided by an animated pedagogical agent concerning syntax and task mistakes or success (e.g., "Check line number 5 of your code. Are you sure that you use the right property?"). Navigation back and forth was possible either stepwise or by jumping to a specific page selectable from a dropdown menu. The learning environment and a sample source code of CSS is shown in Figure 18.

4.2.2.1 Instruments

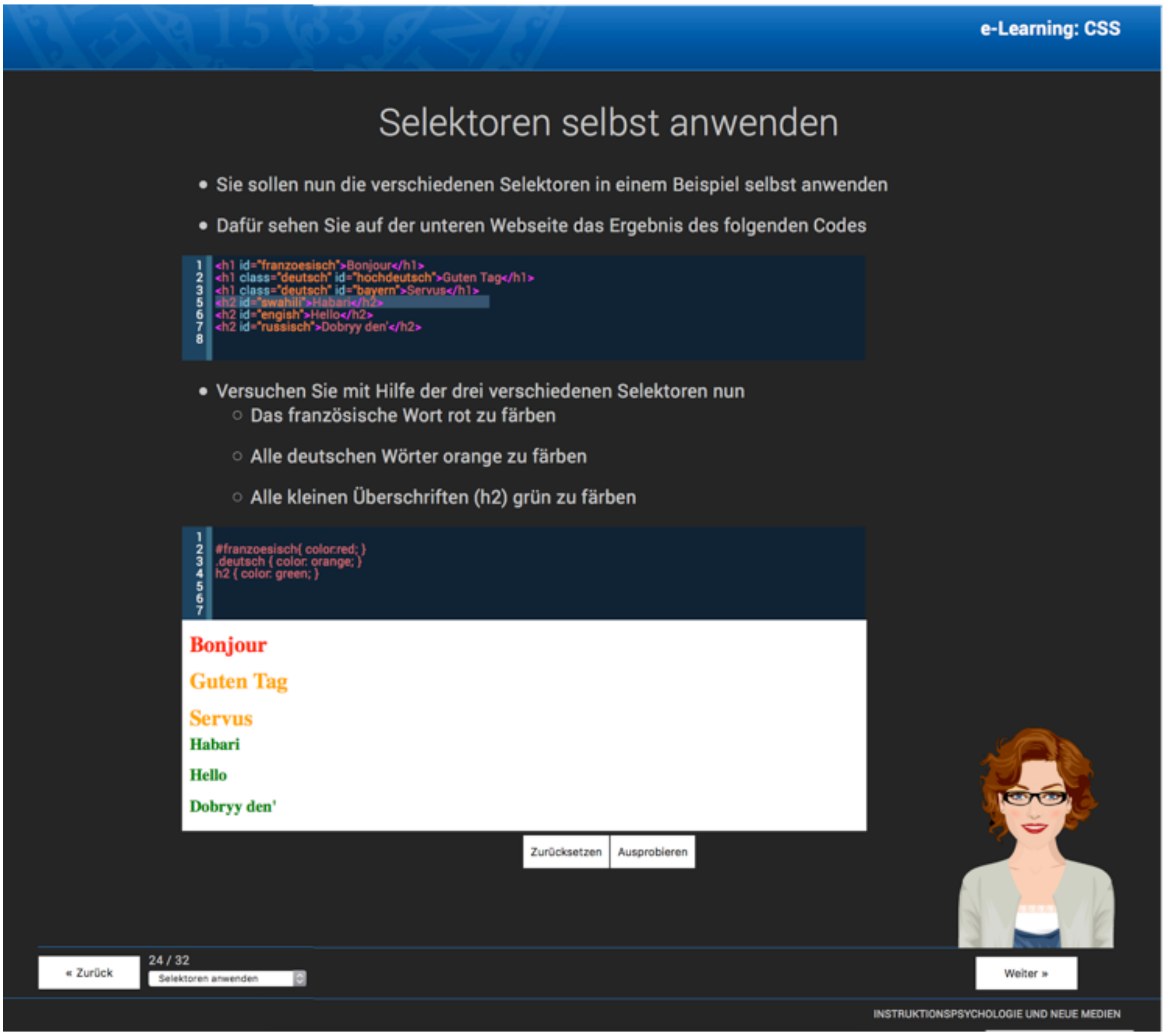
Typing Behavior and Baseline

The JavaScript based ScreenAlytics framework (see chapter 4.1) was implemented into the learning environment to record events triggered by the keyboard. Event data consisted of a timestamp accurate to the nearest millisecond and the pressed keys. Sampling rate for event recognition was approximately 60 times per second.

In order to get a standardized baseline for the typing indices (i.e., typing speed, pauses, corrections), the typing behavior of all participants was recorded while they copied the German sentence "Franz jagt im komplett verwahrlosten Taxi quer durch Bayern," which contains all letters from A to Z. For the coding speed baseline, participants were asked to copy two lines of CSS codes containing all relevant special characters (i.e., {}#;:.="').

Initial Motivation

Prior to learning, initial motivation was measured using the "questionnaire to assess current motivation in learning situations" (QCM; Rheinberg et al., 2001). QCM asks for the degree to which a participant agrees on 18 sentences related to current motivation for an upcoming learning situation (e.g., "This exercise is a real challenge to me."). The questionnaire is a 7-step Likert scale ranging from 1 (does not apply) to 7 (applies). Internal consistency was Cronbach's $\alpha = .853$.



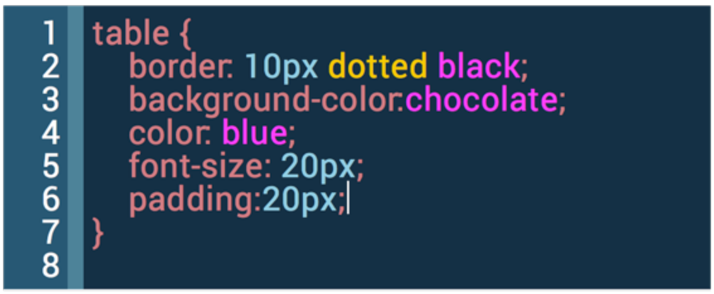


Figure 18. Upper screenshot shows the learning environment used in both studies. Learners typewrite CSS code in a text-area and see the results beneath. Animated pedagogical agent gives feedback regarding mistakes or success. Bottom screenshot shows example CSS source code defining the design of a table element.

Spatial Ability

Spatial ability was assessed as it was found to facilitate learning with multimedia (Münzer, Seufert & Brünken, 2009) and to be relevant in the domain of programming (Jones & Burnett, 2008). It was assessed by using an online version of the VZ-2 paper folding test (Ekstrom et al., 1976), a timed test including 10 problems in which participants have to imagine folding (mentally fold) a square sheet of paper two or three times according to a drawn instruction. In the final instructional drawing, the imaginary folded paper is shown as hole-punched at a specific position. Participants are then required to select the right illustration from five options that shows how the paper would appear when unfolded. The test was timed to three minutes.

Knowledge Tests

Prior declarative knowledge was assessed with 5 single-choice and 12 multiple-choice items (e.g., "Which property changes the font in CSS?"). Internal consistency was Cronbach's α = .889.

Prior procedural knowledge was assessed by an authentic web design coding task. Students were instructed to design a website according to four given design specifications such as "all headings should have a font size of 16px." Codes that were given as answers were rated based on a self-developed rating scale by the author and a research assistant. Interrater reliability was Kappa = .89, $p < .01$. In case of disagreement, raters discussed the final rating.

The same instruments were used to assess post knowledge after learning. Internal consistency of the declarative knowledge test was Cronbach's α = .643. Interrater reliability of procedural knowledge test was Kappa = .93, $p < .01$. Again, raters discussed final rating in case of disagreement.

Recall Task

A recall task was presented after finishing half of the learning content. It prompted learners to describe what they learned about the three methods of including CSS code on a website. They were instructed to name and explain them in their own words. For each of the named include-methods, learners could reach up to three points (naming, explaining the functionality, and naming the syntax). Answers were coded by two raters. Interrater reliability was Kappa = .81, $p < .01$. Raters discussed final rating in case of disagreement.

Current Motivation

A short measure of three seven-step Likert-scaled items (e.g., “I am sure I will find the right solution.”) was presented just before two of the interactive coding exercises and before the recall prompt. This was done to keep changes on motivation between the time of measuring motivation and recording the typing behavior in tasks as small as possible. Due to technical issues, answers of the motivation measure prior to the recall prompt were not saved in the database. Internal consistency was Cronbach’s $\alpha = .788$ for the first exercise and Cronbach’s $\alpha = .766$ for the second exercise.

4.2.2.2 Procedure

Students filled out a consent form to participate in the study. All remaining parts were done in an online environment: demographic variables (sex, age, semester), initial motivation (QCM), spatial ability (paper folding task), baselines for typing and coding, prior declarative and procedural knowledge. Students learned for about 45 minutes. After finishing half of the learning content, students were to work on a free recall task. Current motivation was assessed before each of two coding (programming) exercises and before the recall task. They finished with the post-tests for declarative and procedural knowledge.

4.2.3 Results

4.2.3.1 Statistical Analysis and Computation of Scores

The Type I error rate was set to .05 for all analyses. One-tailed tests were used as directional hypotheses were formulated. Outliers were defined as values greater than the upper quartile plus three times the interquartile-range according to conservative statistical definition (Field, 2009, p. 135). IBM SPSS Statistics 22, PHP statistics library, and R were used to analyze the data.

Due to technical issues, data of five participants were missing for analyses regarding declarative and procedural knowledge, the paper folding task, and initial motivation. Table 2 shows the means and standard deviations of the central variables.

Table 2

Means and standard deviations of important variables.

Variables	Pre		Post	
	M	*SD*	*M*	*SD*
Declarative knowledge (Max = 46)	11.03	10.92	32.42	5.92
Procedural knowledge (Max = 15)	1.24	3.84	11.29	3.69
Spatial ability (Max = 10)	6.54	2.23	-	-
Initial motivation (Max = 7)	4.20	.84	-	-
Current motivation, Exercise 1 (Max = 7)	-	-	4.30	1.54
Current motivation, Exercise 2 (Max = 7)	-	-	3.83	1.60

Note. N = 38

Correctly solved items of pre and post declarative and procedural knowledge test were summed up to individual scores. Item values of the QCM were summed up to a total initial motivation score. Items of the short-scale for current motivation were summed up to total scores separately for exercise 1 and 2.

The following indices / features were extracted for typing behavior: speed as the ratio of keystrokes and typing time, frequency of short pauses, long pauses, total number of keystrokes and deletings. Pauses were defined as not registering keystrokes for 1 to 6 seconds (short pause,]1;6[) and 6 to 60 seconds (long pause, [6;60]). The threshold of 6 seconds represents the rounded value of a median split of all pause durations in typing tasks. Pauses were removed if the duration exceeded 60 seconds or if a page change was recognized between two keystrokes. Individual means of all indices were computed separately and overall for all interactive exercises as well as for the recall task. Standardized means of typing indices for the recall task and for the interactive coding exercises were computed by subtracting the assessed baseline values.

4.2.3.2 Fast-Typing-High-Performance-Hypothesis

Bivariate Bravais-Pearson correlations were computed between the indices of typing behavior that indicate fast typing during the recall task and a) performance in recall task, b) pre/post declarative knowledge, and c) pre/post procedural knowledge. Against the assumptions, at least one indicator of fast typing significantly negatively correlated with performance in recall task, and prior declarative, procedural, and post declarative knowledge. Correlations with post procedural knowledge were not significant. Table 3 shows detailed correlations. Note that higher numbers of short pauses, long pauses, keystrokes, and deletings indicate slower typing.

More than one index of typing behavior was significantly correlated with recall performance and post declarative knowledge. Thus, multiple linear regressions were computed for those variables to identify the variance explained by the typing indices. As strongly correlated predictor variables tend to bias multiple regression models, collinear typing indices were identified and subsequently rejected. A correlation threshold of Pearson's $r > .7$ was adopted to remove them. If two variables were collinear, the predictor with a stronger correlation with the specific performance was kept.

Hence, for recall performance, the number of short pauses and the number of keystrokes were kept. For post declarative knowledge, both typing speed and the number of short pauses were kept. For prior declarative and procedural knowledge only the number of short pauses was significantly correlated.

Table 3

Bravais-Pearson correlations between indices of typing behavior during recall task and performance variables.

	Pre			Post	
	Recall	Declarative	Procedural	Declarative	Procedural
Typing speed	-.129 (p = .233)	-.281 (p = .063)	-.264 (p = .075)	-.376 * (p < .05)	.168 (p = .193)
Number of short pauses	.643 ** (p < .001)	.518 ** (p < .001)	.332* (p < .05)	.360 * (p < .05)	.128 (p = .242)
Number of long pauses	.131 (p = .214)	-.093 (p = .300)	-.197 (p = .132)	-.027 (p = .441)	.080 (p = .332)
Number of keystrokes	.709 ** (p < .001)	.212 (p = .115)	.039 (p = .413)	.020 (p = .456)	-.075 (p = .341)
Number of deletings	.576 ** (p < .01)	.244 (p = .086)	.129 (p = .238)	.054 (p = .383)	.105 (p = .288)

Note. * $p < .05$; ** $p < .01$; $N = 38$

In order to control variability among students, regression models including the following predictors were first computed: prior declarative knowledge, prior procedural knowledge, and spatial ability. The residuals of each regression were then entered into separate secondary regressions, including the previously identified typing indices as predictors. Thereby, the unique variance of the particular typing indices could be determined. Table 4 shows the summaries of the conducted multiple regression models.

The analyses of the typing indices of the recall task discovered statistically significant relationships for recall and prior declarative knowledge. Regarding recall, the typing indices improved prediction of the recall performance by 23.9%. Regarding prior declarative knowledge, only spatial ability was used as a predictor and the first model was not significant. However, the second model was significant with the number of slow pauses predicting 24.5% of the variance. For post declarative knowledge, the non-significance of the second model implies that the two typing indices did not add a significant improvement in prediction to the first model.

Table 4

Summaries of the multiple regression models for performance on recall, prior and post declarative knowledge predicted by typing behavior during the recall task.

Model / Predictors	sig.	df1, df2	Recall		Prior Declarative		Post Declarative	
			R^2_{adj}	F	R^2_{adj}	F	R^2_{adj}	F
(I) DK_{pre}, PK_{pre}, SA	.002	3, 29	0.334	6.35	-	-	-	-
(II) SP, KS	.008	2, 28	0.239	5.71	-	-	-	-
(I) SA	.456 (n.s.)	1, 33	-	-	-.013	0.57	-	-
(II) SP	.002	1, 32	-	-	.245	11.71	-	-
(I) DK_{pre}, PK_{pre}, SA	.004	3, 31	-	-	-	-	0.346	5.470
(II) SP, TS	.649 (n.s)	3, 30	-	-	-	-	0.033	0.469

Note. * p < .05; ** p < .01; N = 38; PK_{pre} = prior procedural knowledge, DK_{pre} = prior declarative knowledge, SA = spatial ability, SP = number of slow pauses, KS = number of keystrokes. First step (I) was done to control for individual differences in PK_{pre}, DK_{pre} and SA.

4.2.3.3 Fast-Coding-High-Performance-Hypothesis

Bravais-Pearson correlations were computed between the same indices of typing behavior and performance and examined the typing behavior during interactive coding exercises. In line with the assumptions, each performance measure was correlated with at least one indicator of fast typing. Table 5 gives an overview of the detailed correlations. Note that higher numbers of short pauses, long pauses, keystrokes and deletings again indicate slower typing.

After that, multiple regression analyses were computed as described in the previous hypothesis. Conducting the collinearity analyses identified the following predictors to keep for further analyses: for recall performance, the typing speed, and number of short pauses were kept. For prior procedural knowledge, only typing speed was kept. For post declarative knowledge and for post procedural knowledge, all variables were kept. Table 6 shows the summaries of the conducted multiple regression models.

Table 5

Bravais-Pearson correlations between indices of typing behavior during interactive coding examples and performance variables.

		Pre		Post	
	Recall	Declarative	Procedural	Declarative	Procedural
Typing speed	.444 * (p < .05)	.288* (p < .05)	.316* (p < .05)	.429 ** (p < .01)	-.046 (p = .400)
Number of short pauses	-.499 ** (p < .01)	-.269 (p = .054)	-.259 (p = .061)	-.469 ** (p < .01)	-.392 ** (p < .05)
Number of long pauses	-.212 (p = .094)	-.179 (p = .141)	-.309* (p = .029)	-.361 * (p < .05)	-.417 * (p < .01)
Number of keystrokes	-.212 (p = .094)	-.170 (p = .157)	-.172 (p = .155)	-.269 (p = .054)	-.536 ** (p < .01)
Number of deletings	-.359 * (p < .05)	-.260 (p = .057)	-.166 (p = .160)	-.364 * (p < .05)	-.251 (p = .073)

Note. * p < .05; ** p < .01; N = 38

Table 6

Summaries of the multiple regression models for performance on recall, post declarative and procedural knowledge predicted by typing behavior during interactive coding exercises.

Predictors	**sig.**	**df1, df2**	**Recall**		**Post Declarative**		**Post Procedural**	
			R^2_{adj}	F	R^2_{adj}	F	R^2_{adj}	F
(I) PK_{pre}, DK_{pre}, SA	.003	3, 28	0.330	6.09	-	-	-	-
(II) SP, TS	.007	2, 29	0.244	6.00	-	-	-	-
(I) PK_{pre}, DK_{pre}, SA	.004	3, 31	-	-	0.283	5.46	-	-
(II) SP, LP, TS	.189 (n.s.)	3, 29	-	-	0.062	1.70	-	-
(I) PK_{pre}, DK_{pre}, SA	.095 (n.s.)	3, 28	-	-	-	-	0.115	2.34
(II) SP, LP, KS	.004	3, 30	-	-	-	-	0.285	5.38

Note. N = 38; PK_{pre} = prior procedural knowledge, DK_{pre} = prior declarative knowledge, SA = spatial ability, SP = number of slow pauses, LP = number of long pauses, KS = number of keystrokes.

Typing indices of interactive coding exercises significantly improved prediction of recall performance by 24.4% and of post procedural knowledge by 28.5%. Typing indices could not add significant improvement in predicting post declarative knowledge.

4.2.3.4 Typing Behavior and Motivation

Relations were explored between typing behavior and motivation by conducting two-tailed bivariate Bravais-Pearson correlations of the typing indices including 1) the recall prompts, 2) all interactive coding exercises, 3) the exercise that followed the first current motivation measure and 4) the exercise that followed the second current motivation measure, with a) the initial motivation, b) the first measure of current motivation and c) the second measure of current motivation. Table 7 shows the correlation coefficients with the motivation variables (a-c) in the first and the task of which the indices were computed in the second line (1-4).

Table 7

Bravais-Pearson correlations between initial motivation and current motivation and typing indices during different tasks.

	a) Initial Motivation				b) Current Motivation, Exercise 1			c) Current Motivation, Exercise 2		
	1) Recall	2) Exc., Overall	3) Exc. 1	4) Exc. 2	2) Exc., Overall	3) Exc. 1	4) Exc. 2	2) Exc., Overall	3) Exc. 1	4) Exc. 2
TS	-	-	-.225 $p = .186$	-	-	-	-	.226 $p = .167$	-	-
SP	.374*	-.304 $p = .064$	-	-.276 $p = .120$	-	-.356*	-	-	-.279 $p = .085$	-
LP	-	-.301 $p = .063$	-	-.244 $p = .165$	-.335**	-.388*	-.249 $p = .142$	-	-.242 $p = .138$	-
DEL	-	-.237 $p = .146$	-.305 $p = .071$	-	-.230 $p = .153$	-.540**	-.309 $p = .067$	-	-.284 $p = .080$	-.227 $p = .176$
KS	.253 $p = .143$	-		-		-.354*	-	-	-	-

Note. * $p < .05$; ** $p < .01$; $N = 38$; Motivation variable is listed in the first line, second line lists from which task the typing indices were computed. All correlation coefficients with p < .200 were reported. TS = Typing speed, SP = Number of slow pauses, LP = Number of long pauses, DEL = Number of deletings, KS =Number of keystrokes.

Analyses revealed the same pattern for motivation as for performance. At least one typing index for recall task correlates significantly positive with initial motivation, $r = .374$. In contrast, only significantly negative correlations were found from $r = -.540$ to $-.354$ between typing indices during exercises and initial / current motivation measures. Actual typing speed (ratio of keystrokes and typing time), however, did not correlate significantly with any motivation measure.

4.2.4 Discussion

4.2.4.1 What Was Done and Found

As discussed in the introduction of this study, it is mostly regarded as common sense that writing processes can tell about cognitive processes. However, the interpretation of writing processes regarding different psychological variables is still very difficult. Thus, this study examined how differentiated indices of typing behavior correspond with performance and motivation in a learning environment about website programming. Two different types of writing tasks were used. On the one hand, learners had to produce open text and summarize methods of using CSS code on a website. On the other hand, learners had to write program code in order to solve given design problems.

It was expected that indices of higher typing speed correspond with higher performance regarding recall and post knowledge. However, the results clearly showed that the opposite was true: subjects who showed a slower typing behavior (i.e., lower typing speed, higher number of short pauses) performed better on the recall task, the prior declarative and procedural knowledge tests, and the post declarative knowledge test. Although the relations with post procedural knowledge were not statistically significant, they tended to the same direction. Thus, the Fast-Typing-High-Performance hypothesis needs to be rejected and revised.

At the same time, it was expected that subjects who show a higher typing speed while working on interactive coding exercises would tend to show higher performances. Interestingly, this was found to be true regarding recall performance, prior, and post knowledge. According to these results, the hypothesis has been confirmed.

4.2.4.2 How to Interpret the Results

Taking the results of the two hypotheses into account, the findings show that typing behavior of continuous text has to be interpreted reverse to typing behavior of typing program code when one wants to link it with task performance. This is counterintuitive but comprehensible: Coding tasks require the recall of previously learned proceduralized chunks of code following a given script comparable to an instruction manual. Learners who are able to do a fast and correct recall will have a higher typing activity. Learners who can correctly code syntax will make less mistakes and therefore show less activity in deleting and correcting their code. In contrast, typing continuous text in the open recall task requires the reconstruction and verbalization of declarative knowledge. Although this needs to be investigated again, learners who show slower overall typing speed might have a higher conscientiousness. Frequent corrections and more pauses seem to indicate a high persistence and the set of a high standard. The different requirements of the tasks are comparable with building constructions from Lego bricks. An expert Lego builder will be fast when following given building instructions but will probably make more corrections and will need more bricks when we ask her to freely replicate a model of her house with Legos.

4.2.4.3 Finding Useful Indices

Some of the indices seemed to be collinear due to their operationalization (e.g., number of keystrokes should be higher when learners make a lot of corrections / deletings). Identifying unique typing indices has been very important in order to not overestimate the relations and will help examining the right features in future studies. Alike, it was important to examine the task performances without the influence of personal characteristics of knowledge and spatial ability. The achieved explained variances of up to 28.5% show the high potential of analyzing typing behavior.

4.2.4.4 Motivation

Regarding the explorative analyses of the association of typing behavior with motivational states, the results seem to follow the pattern of task performance. Learners that show a higher typing speed during the recall task tend to experience lower (initial) motivation whereas higher typing speed during exercises indicates higher (current) motivation. While this was significant for the first exercise that was analyzed, it could not be found clear-cut in the second exercise. Unfortunately, due to technical issues, measurement for current motivation was not available before the recall task.

4.2.4.5 Methodological Challenges

Designing studies about the correspondence of data channels with established measures of constructs brings about methodological challenges. One could argue that a correlation study is not appropriate to examine and understand a new data stream as one cannot draw causality from correlation. However, with this first investigation of the relationship between typing behavior and learner variables, the stage is set for further investigations of causality and deliver important hints on indices of typing to look at. Moreover, considering the main objective to predict variables in order to adapt learning environments, causality is not as essential as it would be for a work that is solely dedicated to deepening our understanding of underlying theoretical assumptions.

Additionally, many of the presented studies in the literature review use machine learning approaches to classify or predict latent variables. Of course, future studies on possible adaptions should investigate the accuracy of machine learning algorithms as well. However, it is argued that it is important to find out about correlations as a first step because machine learning algorithms draw the curtain over underlying mechanisms that help us to understand relations between indices of observable behavior and latent variables (see Kitchin, 2014 for an overview).

Another challenge is that variables are not stable throughout a learning session. In this study, current motivation is a variable that changes during the learning process. Thus, a continuous measurement is needed to model it. Even though typing behavior can only be measured when learners currently work on writing tasks, it is continuous within these tasks. A fundamental question when analyzing continuous data is how long are the segments that we observe for a prediction. In this study, different time segments were not tested but it was looked at the whole task. This needs to be addressed in further investigations.

4.2.4.6 Conclusions & Future Directions

Recording and analyzing peripheral data to predict variables relevant for learning has many advantages compared to other objective measures, such as physiological data, eye-tracking data, or log-files. It has no special hardware-requirements, is unobtrusive, non-reactive, and relatively easy to implement and analyze. However, compared to other measures, it lacks valid examinations of possible correspondences to relevant latent psychological variables. This study attempted to reduce this discrepancy by systematically labelling data of typing behavior with measures of latent state variables which are relevant for learning, namely, motivation and task performance.

The study confirms that there is a relationship between the typing behavior and achieved task performances. However, this typing behavior needs to be interpreted task-specifically.

The results of this study show that it is worth opening the black-box of the more commonly used log-files. While log-files only provide us with information about where and when a learner navigated or about specific pre-defined events, peripheral data gives us a more detailed insight into what happens during the learning process. Given this high resolution and granularity of measures for behavior in online environments, operationalization of latent constructs can be more accurate compared to classic log files.

There is still need for further validation of the presented results regarding different domains and contents. As typing behavior was found to be task specific, it is assumed that within continuous texts it can be interpreted independent from content, but this needs to be investigated with other contents.

More research is also needed to interpret, generalize, and specify this large set of information regarding different latent variables, upcoming and changing peripheral data of mobile devices, and of course, applications to enhance learning.

As recording and analyses of typing behavior can be done at runtime, the prediction of performance and motivation could be applied to improve both timing and content of instructional support. Detecting motivation could be used to improve the adaption of the difficulty of the presented learning content or prompt learners. It is important to note that, although some of the examined indices could explain considerable variance of the latent variables, the presented data stream cannot be used as the only instrument to measure learner variables or adapt interventions. Multimodal data is needed to make appropriate adaptions. Combining a set of predictive measures such as peripheral data together with rapid assessment tasks (Kalyuga, 2008) is a promising approach that should be followed and evaluated in future research.

4.3 Study 2: How Mouse Behavior Corresponds with Cognitive Load and Affect States

Reliable and valid measurement of experienced CL (see chapter 2.1) is a theoretical and methodological issue that has been discussed for decades (e.g., Brünken et al., 2003; Klepsch et al., 2017). Moreover, real-time process measures of CL that are reflective of the dynamic nature of self-regulation and CL are still demanded in recent literature (Seufert, 2018). As high ECL can hinder learning, finding a suitable real-time measure for it would offer a range of possibilities, both for recognizing badly designed environments, but especially for adaptive learning environments that could then adapt the difficulty or level of support to the needs of learners. In this quasi-experimental study, the relationship between mouse behavior and CL is investigated. More precisely, it examines whether pauses in the interaction with the learning environment (no mouse and keyboard use) are associated with increased CL. Moreover, detailed peripheral data as introduced in chapter 4.1 were recorded to perform explorative analyses regarding correlations with affect scales and learning performance.

4.3.1 Measurement of Cognitive Load

Various indicators are used to measure CL. Wierwille & Eggemeier (1993) distinguish between three main categories for measuring cognitive load: physiological, subjective and task- or performance-based indicators. Other authors (Brünken et al., 2003, 2002) classify the available measurement methods by the two dimensions of objectivity (subjective or objective) and causal reference (direct, indirect).

The objectivity dimension describes whether the method records subjective, self-reported data, or objective observations of sources such as behavior (e.g., reaction times), physiological reactions (e.g., heart rate) or learning outcomes. The dimension of the causal reference classifies the methods according to whether the observed phenomenon has a direct or indirect relation to CL. For example, there is a direct relationship between CL and the self-reported difficulty of learning materials, because difficulty is directly related to intrinsic and extraneous load. An indirect relation results, for example, between measures of the learning outcomes and the CL, because the theory assumes that the learning performance decreases due to a high CL (Brünken et al., 2002). Examples of physiological indicators are heart rate (Paas & van Merriënboer, 1994), pupil dilation (Beatty, 1982; Van Gerven, Paas, Van

Merriënboer & Schmidt, 2004) or EEG (Antonenko, Paas, Grabner & van Gog, 2010). These measure CL indirectly, for example, high CL could lead to an increased heart rate. However, it is also possible that e.g., the emotional reaction to the learning material is responsible for these changes (Brünken et al., 2003).

Subjective indicators work with a self-report of learners regarding their CL during or after learning. For example, an indirect measurement can be the subjectively reported level of mental effort that a learner puts into the understanding of learning materials (e.g., Paas, van Merriënboer & Adam, 1994). Such self-report techniques are widely used in CL research (Paas, Tuovinen, Tabbers & van Gerven, 2003). However, researchers doubt the ability of individuals to rate their load with high accuracy (e.g., Schnotz & Kürschner, 2007). More recent self-report measures are able to measure differentiated types of CL (Klepsch et al., 2017), either by informing learners about CLT before letting them report about their CLT or by using a naïve rating without such training. A second subjective measurement used by Kalyuga, Chandler and Sweller (1998), for example, allows the persons to rate how difficult the learning material is. As mentioned, this self-reported difficulty refers directly to the CL. Kalyuga et al. (1998) reported a high sensitivity of these scales for differences in the preparation of training. Brünken et al. (2003) criticize, however, that these differences can also be explained by individual competence levels or different levels of attention. As CL is dynamically changing during the learning process, depending on the learning material and the cognitive constitution of the learner, self-report measures face the inherent drawback that ratings cannot be acquired during the cognitive action of interest (e.g., Schmeck, Opfermann, van Gog, Paas & Leutner, 2015).

The paradigm of dual task provides an objective, direct and online measure. It is based on the assumption that the limited cognitive resources can be shared flexible among parallel tasks. Simultaneously to a primary task (usually a learning task), an artificial secondary task is presented. The performance of the secondary task is directly associated with CL: if the primary task needs a high amount of cognitive capacity, performance of the secondary will decrease. Therefore, it is necessary to assess a baseline of the solely execution of the secondary task without being loaded by the primary. An often used secondary task consists of the learner monitoring an element and reacting to its changing color. Reaction time then indicates the amount of CL (Schoor, Bannert & Brünken, 2012). Another example is the execution of an internalized task such as food tapping a previously practiced rhythm. Precision of the executed rhythm serves as an indicator for CL (Park & Brünken, 2015). In working memory research, the dual-task method has long been the first choice (e.g., Baddeley & Logie, 1999). Surprisingly, this method has long been neglected in CL research and multimedia learning (Brünken et al., 2002; Chandler & Sweller, 1996; Marcus, Cooper, & Sweller 1996; Sweller, 1988). It offers a promising tool for the direct measurement of ECL (Brünken et al., 2002) and is therefore also used in this study.

Taking this into account while considering navigation as a discrete task that requires cognitive resources in the learning process and operationalizing navigation as different indices of mouse behavior, one goal of this study is to reveal a relationship between mouse behavior and CL.

Mouse behavior and CL

Measuring CL through mouse behavior is not an entirely new idea. A first approach of relating mouse movements and CL was done by Arshad, Wang and Cheng (2013). Participants were presented environments that induce high and low CL in a simulated computer-based platform to screen applicants for a fictitious human resource department. The authors found a higher frequency of pauses in mouse movements for the high CL environment. However, there was no additional measurement of the CL aligned to the mouse behavior. Thus, the individually experienced CL by the learners could not be controlled.

In another study, Grimes and Valacich (2015) examined the relationship between fine-motor control, operationalized through indices of mouse behavior, and CL. The authors found significant differences of some indices of the mouse behavior (Euclidean distance and slow movements) between three tasks with different difficulty. However, the used materials were very artificial, asking the participant to verify viewed numbers on an otherwise blank screen. In a similar artificial task, Rheem, Verma and Becker (2018) showed that slower movements and less trajectory deviations corresponds to a higher CL. It is worth noting that once again the study did not validate whether CL was imposed on participants as intended.

4.3.2 Affective State

Regarding affective states, a lot of research has been done trying to assess them by using psycho-physiological sensors (e.g., Hudlicka & McNeese, 2002; Rani, Sarkar & Smith, 2003), facial features (e.g., Cohn & Kanade, 2006), vocal features (Banse & Scherer, 1996; Batliner, Steidl, Hacker & Nöth, 2008; Cowie et al., 2001), and linguistic or conversational features (D'Mello et al., 2008; D'Mello, Craig, Witherspoon, McDaniel & Graesser, 2007; Vizer, Zhou & Sears, 2009). As D'Mello, Craig, Witherspoon, McDaniel and Graesser (2007) state, using obtrusive measures (e.g., physiological sensors) to predict affective states would distract the learner and interfere with the primary task. In contrast, mouse behavior is no artificially added task and therefore is unobtrusive and non-reactive as a potential measure for affective states. Moreover, facial features need special hardware, and vocal, linguistic and conversational features need learners to speak loudly or perform writing tasks. Although some research exists that links mouse and keyboard data to affective states (Kolakowska, 2013), hardly any attempts have been made in educational research. Existing studies try to detect and classify binarily whether learners show a specific academic emotion or not (e.g., enjoy / not enjoy in Lali, Naghizadeh, Nasrollahi, Moradi & Mirian, 2014; bored / not bored in Tsoulouhas, Georgiou & Karakos, 2011), but affective states have not yet been investigated.

Thus, learners' affective state are examined from the perspective of a common two-dimensional model that subsumes affect and valence in orthogonal factors of positive and negative affect (Watson & Tellegen, 1985). The used instrument constructs positive and negative affect as distinct interval-scaled measures. Watson, Clark & Tellegen (1988, p. 1083) summarize positive affect (PA) as "the extent to which a person feels enthusiastic, active, and alert. High PA is a state of high energy, full concentration, and pleasurable engagement, whereas low PA is characterized by sadness and lethargy". In contrast, they describe negative affect (NA) as "a general dimension of subjective distress and unpleasurable engagement that subsumes a variety of aversive mood states, including anger, contempt, disgust, guilt, fear, and nervousness, with low NA being a state of calmness and serenity" (p. 1083). As activation is inherent to the level of both dimensions of experienced affect, it is argued that the level of activity in using peripheral devices should correspond with affective state. For instance, higher mouse speed or more frequent mouse movements should go along with higher positive or negative affect. Like Yannakakis, Hallam and Lund (2008) claims for psycho-physiological measures, it is argued that by analyzing the mouse behavior one cannot distinguish negative (e.g., anger) from positive affect (e.g., pleasurable excitement), but only the level of activation.

4.3.3 Research Question and Hypotheses

In this study, navigation is operationalized as different indices of mouse behavior. Moreover, navigation is considered to be a discrete task that represents an ECL while learning. Those assumptions lead towards the question whether mouse behavior changes depending on experienced CL and therefor can be used as an information source of the CL experienced by the learner. This study wants to account for the fact that currently available studies did not apply aligned validation measures of CL. Hence, this study investigates the correspondence of mouse behavior with an established online dual-task reaction-time measure.

It is hypothesized that increased CL leads to pauses in the mouse behavior, as there are not enough resources available to spend on this task. Respectively, such pauses indicate increased CL (*Hypothesis 1: No-Interaction-High-Load-Hypothesis*).

In the second research question, it is asked whether it is possible to draw conclusions about affective states by mining peripheral data. Therefore, the associations of mouse behaviors with affect is explored. It is hypothesized that higher activity in mouse behaviors correspond with higher positive and negative affect (*Hypothesis 2: Active-Mouse-High-Affect-Hypothesis*). As mouse activity appears in more than one feature, the relationship is checked between learners' affective states and the following indices: mouse speed, covered distance, number of short pauses in mouse behavior, number of mouse clicks, number of scrolling activities. It is argued that the number of short pauses represents a higher mouse activity because it indicates the frequency of initialized movements. Hence, the more movements learners start, the higher should be their level of affective states.

4.3.4 Method

4.3.4.1 Sample and Design

In a quasi-experimental study, $N = 49$ undergraduate students majoring in media communication at a German university participated and learned about website programming. All were enrolled in one of three parallel courses ($n_1 = 21$; $n_2 = 16$; $n_3 = 12$) dealing with the basics of media production such as image editing or web design. Participants received no incentives but learning contents were part of the course curriculum. Students in course 2 and 3 were assigned to the experimental group, students in course 1 to a control group. Due to the quasi-experimental design, more students were assigned to the experimental group ($n_{EG} = 28$; 4 male; $M_{age} = 20.11$; $SD = 1.39$) than to the control group ($n_{CG} = 21$; 3 male; $M_{age} = 20.48$; $SD = 1.63$).

In order to compare CL during pause-situations (not using mouse and keyboard) with interaction-situations (using mouse or keyboard), the timing of dual-task CL assessments was manipulated. In the experimental group, CL measurement was triggered after a random time interval of 2 to 10 seconds which started counting down only if learners did not interact with their mouse and keyboard for more than 6 seconds. In the control group, measurement was triggered after random time intervals of 15 to 35 seconds irrespectively of their peripheral device usage. Intervals were based on the mean pause times and mean visiting times of a previously conducted pilot study in which the learning materials were tested. A between-subject design was used instead of a within-subject design to reduce testing frequency in experimental group and ensure that learners in the control group cannot influence the measurement timings.

The laboratory was equipped with 21 iMac desktop computers (21,5 inch display with a resolution of 1920x1080, tethered apple mouse and keyboard). Firefox was used as web browser and learning environment was presented in full screen mode.

4.3.4.2 Learning Materials

Students had to learn the basic syntax and properties of “cascading style sheets” (CSS), a common standard to style websites. Learning material of the first study was used, but minor changes were applied to the content (see chapter 4.2.2.2). It was structured linear and consisted of 15 content pages including about 1700 words, two tables, one quiz, 12 code examples and three interactive exercises. In interactive exercises, learners had to write CSS code to solve a given task (e.g., “Set the width of the image to 200 pixels”). The results of their code were presented below the text area when clicking on a “Try it!” button as well as verbal feedback was provided by an animated pedagogical agent concerning syntax and task mistakes or success (e.g., “Check line number 5 of your code. Are you sure that you use the right property?”). Navigation back and forth was possible either stepwise or by jumping to a specific page selectable from a dropdown menu.

4.3.4.3 Measures and Instruments

Peripheral data

The ScreenAlytics software framework (see chapter 4.1) was implemented into the learning environment to record events triggered by the mouse and keyboard. To reduce server load, the events were first recorded into a client-side array. Every five seconds or when leaving a website, the data was sent to a database server and cleared on client side. Event data consisted of a timestamp accurate to the nearest millisecond, the type of event and specific details such as x/y position of the mouse, scroll position or the pressed key. The following event types were recorded: mouse move, mouse click, scroll, keystroke down, window resize / website zooming. Sampling rate for event recognition was 60 times per second.

Cognitive load

Cognitive load was measured by using the dual-task approach (e.g., Schoor, Bannert, & Brünken, 2012). During the primary learning task, the subjects were instructed to monitor the website’s background color as a spatially contiguous secondary task. Subjects were instructed to react to changings of the background color from black to red as fast as possible by pressing the ESC key on their keyboards. Reaction times between color changings and keystrokes were measured as an indicator for CL. Background color was set back to black when pressing the ESC key. To standardize the measures, an individual baseline was assessed prior to learning.

Knowledge tests

As learners were expected to be novices in website programming, prior knowledge was checked with the five-step Likert-scaled item “How well can you write CSS code?”. For post-test, declarative knowledge was assessed with 5 single choice items (e.g., “Which property changes the font in CSS?"). The reliability of the scale was Cronbach's $\alpha = .33$. Procedural knowledge was assessed by an authentic web design task. Subjects were instructed to design a website according to four given design specifications such as "all headings should have a font size of 16px.". Answers were rated based on a self-developed rating scale by the first author and a research assistant. Interrater reliability was *Kappa* = .89, $p < .01$. In case of disagreement, raters discussed the final rating.

Positive and negative affect

Affective states of the learners were measured prior and after learning with the German version of the Positive and Negative Affect Schedule (PANAS, Krohne, Egloff, Kohlmann & Tausch, 1996; Watson et al., 1988). PANAS asks for the degree to which participants experience 20 different feelings related to positive affect and negative affect, using a slider ranging from 1 (not at all) to 100 (very much). It is an established measure of affect and has been successfully used in a range of experiments dealing with affects in learning (e.g., Plass, Heidig, Hayward, Homer & Um, 2014; Um, Plass, Hayward & Homer, 2012). Separate individual scores for positive and negative affect were obtained by computing the mean of each scale. Reliabilities for pre/post, positive/negative affect scales were between Cronbach's $\alpha =$ .88 - .91.

4.3.4.4 Procedure

Students filled out a paper consent form to participate in the experiment. All remaining parts were done in an online environment. Subjects were instructed to keep their left hand close to the ESC key while learning. Demographic variables (sex, age, semester) and prior knowledge were assessed followed by five baseline measurements of CL and filling the PANAS scale. After that, students learned for about 30 minutes, then attended PANAS again and finished with post-tests for declarative and procedural learning performance.

4.3.5 Results

4.3.5.1 Statistical Analysis and Computation of Scores

The Type I error rate was set to .05 for all analyses. One-tailed tests were used as directional hypotheses were formulated. Outliers were defined as values greater than the upper quartile plus 3 times the interquartile-range according to conservative statistical definition (Field, 2009, p. 135). IBM Statistics 22, PHP statistics library and R were used to analyze the data. Table 8 shows the means and standards deviation of the central variables in both conditions.

Regarding CL, individual baselines were computed including five measures prior to learning. Individual mean CL was computed for each page by subtracting the baseline from the mean reaction time. Baseline did not differ significantly between experimental group (M_{EG} = 476.89, SD = 61.72) and control group (M_{CG} = 477.95, SD = 55.36), $t(46) = -.063$, $p = .950$. On three learning content pages, CL was not measured at least one time for every participant because they either did not stay long enough on the page to trigger a measure (control group) or did not pause their interaction long enough to trigger a measure (experimental group). Mean CL was computed over 12 pages that all included measures of more than 20 subjects per group. Three pages included measures of less than 20 subjects per group. Cognitive load values of two subjects of the experimental group were removed because of being outliers according to the definition above.

Correctly solved items of post declarative knowledge test were summed up. Self-reported prior knowledge varied between 0 and 3 (M=.84; SD=.94). Influence of self-reported prior knowledge on dependent variables was checked. Subjects with prior knowledge did neither significantly differ in experienced CL, $t(45) = .592$, $p = .557$, nor in declarative learning outcome or any of the affect measures. Post procedural knowledge did differ between learners without (*M* = 6.99, *SD* = 3.90) and with (*M* = 9.62, *SD* = 2.50) prior knowledge, $t(47) = -2.89$, $p < 0.01$ with a high effect size of $d = .827$.

Table 8

Comparison of means for important variables in both conditions.

	CL measures randomly timed (control group, *N* = 21)		CL measures only during interaction pauses (experimental group, *N* = 28)				
	M	*SD*	*M*	*SD*	*t*	*p*	*d*
Pre							
Negative Affect (Max = 100)	17.52	16.50	18.37	18.34	-0.167	.868	0.048
Positive Affect (Max = 100)	44.13	14.25	40.19	12.57	1.025	.311	-0.296
Prior knowledge self-report (Max = 4)	1.05	0.92	0.68	0.95	1.368	.178	-0.395
During							
Cognitive Load [ms]	324.18	116.23	426.73	203.84	-2.050	.046	0.596
Mouse speed	.24	.08	.26	.08	-1.027	.310	0.250
Mouse short pauses, mean [s]	2428.70	153.62	2417.84	217.93	0.195	.839	-0.056
Mouse short pauses, frequency	120	47.74	103.86	35.39	1.361	.180	-0.393
Mouse long pauses, mean [s]	15492.30	4254.24	16899.73	3693.80	-1.237	.222	0.357
Mouse long pauses, frequency	36	11.69	39.64	10.02	-1.173	.247	0.338
Post							
Negative Affect (Max = 100)	13.63	17.25	16.87	20.22	-0.591	.557	0.170
Positive Affect (Max = 100)	45.03	20.46	40.79	12.09	0.908	.369	-0.266
Declarative knowledge (Max = 5)	4.10	1.02	4.00	.94	0.350	.728	-0.149
Procedural knowledge (Max = 16)	8.82	2.75	8.03	3.99	0.775	.442	-0.225

The following features were extracted for mouse interaction: speed as covered distance per moving time, covered distance as pixels, frequency and duration of short and long pauses. Pauses were defined as not registering mouse movement, clicking and scrolling for 1 to 6 seconds (short pause,]1;6[) and 6 to 60 seconds (long pause, [6;60]). As there are hardly any experiences with classifying pauses reported in literature, the threshold value of 6 seconds represents the median of pause duration over all content pages. Individual means of features were computed for each page and overall.

4.3.5.2 No-Interaction-High-Load-Hypothesis

The normal distribution of CL was verified using the Kolmogorov-Smirnov test. A *t*-test for independent samples was computed to compare mean CL between control group (M_{CG} = 324ms; *SD* = 116ms) and experimental group (M_{EG} = 426ms; *SD* = 203ms). This difference was significant, $t\ (45) = -2.05$, $p < .05$ with a medium-sized effect, $d = .60$, confirming the *No-Interaction-High-Load* hypothesis. Between-group differences were not significant for mouse behavior indices as well as all affect and knowledge scores. Level of experienced CL correlated significantly negative with post declarative knowledge ($r = -.258$, $p < .05$). Correlation with post procedural knowledge was not significant ($r = -.167$, $p = .13$).

4.3.5.3 Active-Mouse-High-Affect-Hypothesis

To check the second hypothesis, bivariate correlations were computed between pre / post positive / negative affect and the described mouse features. Separate means of the mouse features were computed for the first two and for the last two learning content pages. Those pages represent mouse behavior that is timed closely to the pre and post affect measures. Correlations with pre-affect measures were conducted with the features of the first two pages, correlations with the post affect measures were related to the two last pages. This decreases the influence of the potential bias that affective states are dynamic and change during complex learning (D’Mello & Graesser, 2012). Separate correlations for experimental and control group were conducted. Although statistical power was decreasing by this, possible between-group differences could be controlled.

Detailed correlations are reported in Table 9. The analyses revealed that, in the control group, the covered distance and the number of clicks were significantly positive correlated with positive affect. Mouse speed, covered distance and number of short pauses were correlated significantly positive with negative affect. In the experimental group, number of scrollings was significantly positive correlated with negative affect. However, no correlation with positive affect could be identified. Hence, second hypothesis was partially confirmed.

Table 9

Bravais-Pearson correlations between mouse indices / typing speed and pre / post affective states and post knowledge for experimental and control group.

	Pre				Post							
	Neg. Affect		Pos. Affect		Neg. Affect		Pos. Affect		Decl. Knowl.		Proc. Knowl.	
	EG	CG	EG	CG	EG	CG	EG	CG	EG	CG	EG	CG
Mouse speed	-	-	-.074 p = .354	.361 p = .054	.057 p = .388	.581**	-	-	-	-	-	-
Covered distance	-	-	-.073 p = .356	.508**	.194 p = .162	.664**	-	-	-.293 p = .065	-.024 p = .460	-	-
Number of clicks	.114 p = .282	-.345 p = .063	.086 p = .331	.410*	.393*	-.043 p = .426	-	-	-.508**	.085 p = .361	-	-
Number of scrollings	.376*	.039 p=.433	-	-	-	-	-	-	-	-	-	-

Note. * $p < .05$; ** $p < .01$; EG = experimental group; CG = control group; $N_{EG} = 28$; $N_{CG} = 21$; Correlation coefficients of both groups are reported if there was at least one coefficient with $p < .100$.

4.3.6 Discussion

According to the first research question, it was expected that pauses in the interaction with the learning environment indicate increased CL. As assumed, students in the experimental group - of whom CL was measured only when they did not interact with the learning environment - showed significantly higher CL with a medium effect size compared to the control group, where CL was measured irrespectively of interaction. The chosen research design could have led to an underestimation of the effect as CL was measured during both interaction and pauses in the control group. Hence, CL values during pauses were also included in the control-group mean value. Internal consistency of post declarative knowledge was low, so interpretations on declarative knowledge are limited and potential relations might not have been revealed. Moreover, measuring CL by the dual-task method could have been reactive to itself, learning outcomes and affective states.

Although overall CL is related slightly negative with post knowledge in the data, it cannot be determined by the presented method, whether the increased CL during pauses is productive (germane or intrinsic) or extraneous. What can be argued from both theory and data is that learners might have experienced not only an increased CL, but a cognitive overload while pausing the interaction with the learning environment. Learning the contents might have taken so many resources that no more were available for controlling the mouse movements. The collected data supports this perspective as it indicates that higher CL correlates slightly negative with post declarative knowledge.

These findings are in line with previous empirical investigations of Arshad, Wang, and Cheng (2013), but are different in the aspect that this study used an additionally aligned validation measure of CL, operationalized as reaction times of a secondary reaction-time task. The absence of this validation measure has been a methodological limitation in all previous studies that were found during the documentation of this study (Arshad et al., 2013; Grimes & Valacich, 2015; Rheem et al., 2018).

Connecting these findings to the underlying idea of this work to subsequently use peripheral data as a source in adaptive learning environments, a first suggestion is to time instructional support according to the learners' mouse behavior. Long pauses in the interaction with the learning environment (in this experiment, pauses above 6s which represented the median of pause duration over all content pages) seem to indicate increased cognitive load, so the learner might need instructional support such as presenting cognitive scaffolds or prompts.

Although counterintuitive, this support should not be provided during the pauses because it is unknown whether a learner experience productive or irrelevant load in a specific pause. Presenting support while experiencing ICL could then negatively affect the learning. Instead, an option could be to provide support on pure navigation events, e.g., when a learner changes to the next page. This is in line with empirical evidence on feedback that suggests not to present feedback during the learners' work on solutions (e.g., Narciss & Huth, 2004). Moreover, adaptive learning environments could also adapt the level of difficulty of the learning environment. As an example, in the used learning environment, high or low CL, indicated through mouse behavior, could trigger a different, more appropriate choice of programming tasks, being less or more difficult according to the experienced CL.

Besides using mouse behaviors as a source for adaption in learning environments, it could be applied as an unobtrusive and non-reactive measure for CL in research. It would be a huge advantage to reveal mouse behavior as a valid, unobtrusive, non-reactive instrument without any special hardware requirements. However, this study only states that CL is higher while learners do not interact with the environment. It did not examine the capability to predict the exact level of CL through indices of mouse behavior. Hence, the promising results raise a bunch of new questions that need to be addressed in future studies: what exact indices could predict the level of CL (e.g., length of pauses, frequency of pauses, mouse speed)? How can we distinguish increased load from disengagement? How can we get adequate cut-off values for pauses? Do we need baselines for mouse movements and what could they look like?

Regarding the second research question, peripheral data was expected to correspond with the level of affective states. Relations were examined of mouse behaviors with affective states of learners and expected higher mouse activity to be linked with higher affective states. Findings revealed both significant correlations of mouse indices with positive affect (for mouse speed, covered distance and number of clicks) and with negative affect (for number of short pauses and scrollings). However, no significant correlations between mouse indices and positive affect were found in the experimental group. One reason could be a generally lower positive affect in the experimental group with a medium-sized effect, $d = .27$, that made it more difficult the reveal a relation. As the picture is inconsistent to some extent, further research needs to confirm the links between affective states without experimental variation. Additionally, separated operationalizations for the activity and valence dimensions of affective states would help to clearly identify the correct relations between mouse activity and learner's affective states.

4.4 Study 3: Recognizing Confusion and Item Difficulty Through Mouse Behavior

Research on the detection and measurement of confusion currently lacks a method that is applicable in online learning environments outside the lab. This study investigates the possibility to detect confusion through mouse behavior while answering multi-item scales. This task was chosen as it is relevant for learning in advanced learning environments, but always occurs in a similar structure of questions and answer options. Additionally, correspondences between indices of mouse behavior and the subjective and objective difficulty of items are explored. Finally, the mouse interaction with feeling-of-knowing ratings are investigated as a potentially unobtrusive measure of metacognitive judgements.

What is Confusion and Why Should We Measure It?

According to Pekrun (2016), epistemic emotions are a subset of academic emotions (Pekrun, 2006) that occur as a result of cognitive information processing during a learning process. Confusion can be seen as one specific epistemic emotion (e.g., D'Mello & Graesser, 2012) that is central to complex learning activities (e.g., Pekrun & Stephens, 2016). Although there is a debate on the different theoretical categorization as an academic, epistemic, or knowledge emotion versus an affective state, the general consensus is that confusion is important for learning. According to D'Mello and Graesser (2014), confusion occurs when learners come across incongruences like "impasses, anomalies, contradictions, disruptions of goals, extreme novelty that cannot be comprehended, and interruptions of organized sequences of actions" (D'Mello & Graesser, 2014, p. 290). In general, it is hypothesized to be triggered by an appraisal of incoming information which does not match existing knowledge (e.g., Silvia, 2010). As such, if confusion cannot be resolved, it can result in frustration and boredom, leading to a negative impact on the learning outcome (D'Mello & Graesser, 2012). Such negative experiences can contribute to learners giving up on a learning session (e.g., Baker, D'Mello, Rodrigo & Graesser, 2010). However, it can also be beneficial for learning if the material causes a cognitive disequilibrium of the learner (in a Piagetian sense) and, most importantly, the learner is able to resolve this disequilibrium through deeper cognitive engagement (D'Mello, Lehman, Pekrun & Graesser, 2014). Thus, confusion can also be an intended instructional element in the sense of a desirable difficulty, if used carefully (e.g., Bjork & Bjork, 2011). It is crucial to understand that, at the moment in which confusion occurs, it is not possible to characterize it as constructive or nonconstructive. Depending on how the learner handles the confusion, it can lead to positive or negative effects on learning. This thought led to the concept of a *zone of optimal confusion.* Within that zone, learners' confusion is high enough to foster deeper cognitive engagement but low enough to be resolved by the learner. Graesser (2011) argues that the design of learning materials should aim at reaching this zone in order to facilitate learning processes.

Given the importance of confusion during learning, it is crucial to find valid measures for it in order to 1) identify unintendedly confusing content in learning environments that can then be revised by the authors, 2) control for confusion in experimental settings that examine confusion as an instructional intervention in the context of desirable difficulties, and 3) use confusion as a source for adaptive, technology enhanced learning environments, e.g., in order to support learners in the correct moment with a scaffold they can benefit from. Valid measures that are available in real-time are especially important for the latter mentioned application in adaptive learning environments at distance, as instructors are not physically available to monitor learning processes and regulate learners if necessary. Timely adapted interventions could help learners to resolve confusion before leading to negative consequences.

How is Confusion Currently Measured?

A range of studies exists that seek to measure or detect confusion in technology enhanced learning processes through different data sources. In order to present an overview of the current state of the art, studies listed in a recent review (Arguel, Lockyer, Lipp, Lodge & Kennedy, 2017) were taken that explicitly address confusion as a dependent variable. This list is complemented with studies that use a method or data channel that has not been mentioned in the review. Table 10 lists the mentioned sources and studies that examined it. Data sources for measuring confusion can be broadly categorized into self-reports, behavioral responses and physiological responses. Self-reports often ask learners to rate whether they are confused or not (yes/no), or to rate their level of confusion using Likert-scales (e.g., from 1 to 6). Less common are self-reports that ask learners to verbally express their confusion (among other emotions) during learning, or after learning using video-cued recalls (Sullins & Graesser, 2014). These result in so-called *emote-aloud* protocols (e.g., Craig, D'Mello, Witherspoon & Graesser, 2008) that are currently analyzed by manually coding emotional states. Although not done yet, coding could potentially be facilitated by methods of natural language processing and thereby solving the problem of data not being available in real-time for adaptions. Another self-report approach that has been investigated in order to label interaction data with confusion states is the provision of a "I am confused" button in the interface that learners can press whenever they experience confusion (Conati, Hoque, Toker & Steichen, 2013).

Different data sources of behavioral responses have been examined regarding their suitability to uncover variance of confusion while learning. Facial expressions have a long tradition in emotion recognition using the facial action coding system (FACS, Ekman & Rosenberg, 2005) to break down the facial expression in smaller action units. Initially developed to detect basic emotions (happiness, sadness, surprise, disgust, anger, and fear), it was extended to some academic emotions. For confusion, the action units 4 (lowered brow) and 7 (tightened lids) were identified to be relevant (Craig et al., 2008). These action units can be detected by recording a video of the learners' face and observe the occurrences either manually (e.g., Sullins & Graesser, 2014) or through automated image processing (e.g., with CERT, the computer expression recognition toolbox by Littlewort et al., 2011 as done in Postma-Nilsenová, Postma & Tates, 2015). Facial expressions can also be captured through facial electromyography (EMG) that records muscle activity. Using this method, right and left corrugator supercilii, and right depressor anguli oris were identified to be relevant for confusion (e.g., Durso, Geldbach & Corballis, 2012).

Besides facial expressions, body movements and posture have also been discussed to measure confusion. For example, gross body movement was recorded through chair sensors and the recorded movements were related to human judges of facial expression and self-reported confusion using machine learning classifiers (D'Mello & Graesser, 2007). However, the reported accuracy (*kappa* = .11) is considered to be poor (*kappa* > .20 is considered as fair, *kappa* > .41 as moderate, Landis & Koch, 1977).

If learners interact verbally with the learning environment based on written text or voice input (e.g., in conversational intelligent tutoring agents), the verbal interaction is another source that potentially provides information about confusion. This can be done by analyzing explicit statements of the learner that are interpreted to express confusion (e.g., "I'm confused!" or "Why didn't it work?"). Nonlinguistic features (e.g., "huh?") were also examined but were only related to surprise and not to confusion (Baker et al., 2010). Moreover, feature sets consisting of meta information about the dialog (e.g., number of positive feedback given by the tutoring system) were analyzed and found to be predictive of confusion measured through human judges of facial expression with *kappa* = .26 (D'Mello et al., 2007).

Data coming from eye tracking was also investigated regarding its possible contribution to the measurement of confusion. As confusion is an expression of a cognitive disequilibrium that is reflected in changes on how learners explore presented information (Graesser, Lu, Olde, Cooper-Pye & Whitten, 2005), these changes can be a potential proxy for confusion. Moreover, some positive correlations were found between eye movement patterns (e.g., number and duration of fixations) and subjective measures of confusion (De Lucia, Preddy, Derby, Tharanathan & Putrevu, 2014). However, both eye tracking approaches seem to be highly dependent on the given context and are therefore not easily transferable to other learning environments and materials.

In technology enhanced learning environments, data about the interaction between the learner and the learning environment can always be automatically collected. This can be any description of interactions from simple access log data in a web-based environment, peripheral data with higher granularity (as described in chapter 4.1), to complex traces in virtual reality environments. Sequences and patterns of events in this data can then be interpreted as behaviors or states in the learning process (e.g., task completion or achievements) which researchers need to empirically link to an existing measure of confusion that is regarded as valid (e.g., a self-report). The rationale behind this is that once a pattern that relates to confusion has been identified, the self-report is no longer needed (e.g., Pardos, Baker, Pedro, Gowda & Gowda, 2014).

Besides behavioral responses that can tell about confusion, physiological responses have been investigated. The rationale behind these measures is that interrupting a sequence of action can lead to physiological reactions like changes in heart rate or pupil size (Macdowell & Mandler, 1989). Electrodermal activity (EDA), also known as galvanic skin response (GSR) measures changes of the electric conductivity of the skin through electrodes, usually placed at the fingers or at the palms of the hand. EDA was found to be suitable for detecting high arousal but is not suitable for discriminating the valence of emotions (e.g., van Dooren, de Vries & Janssen, 2012). However, some links between specific events in learning and patterns of the EDA signal were identified. In one study, facing a very difficult problem-solving task that can be interpreted as confusion was found to be reliably linked to a drop in skin conductance (Pecchinenda & Smith, 1996). In a more recent study, a support vector machine (SVM) model including blood volume pressure, heart rate and skin conductance was found to discriminate the correct (self-reported) emotion out of four reported emotions (boredom, confusion, hopefulness, and engagement) with a rate higher than chance (68,1%, Shen, Wang & Shen, 2009).

Although not applicable in actual learning situations, brain imaging (e.g., fMRI) was also discussed as a method to measure confusion. Increased activity of the posterior medial frontal cortex was found when learners experienced unexpected feedback that should result in confusion (Hester, Barre, Murphy, Silk & Mattingley, 2008). As a less invasive and cheaper method with better mobility, electroencephalogram (EEG) was used to detect whether students are confused while watching videos in a MOOC. Performance of the classifier to detect students self-reported confusion was just above chance, but as efficient as human observers that were asked to detect confusion by monitoring the body language of students (Wang et al., 2013).

Table 10

Papers that examined data sources and methods used to measure and detect confusion.

Data / Method	Paper
Self-report	
Binary Choice	Lehmann et al., 2012
Likert-Scales	D'Mello & Graesser, 2014
Emote-Aloud While Learning	Baker et al., 2010
Emote-Aloud Retrospective	D'Mello & Graesser, 2014
Explicit reports in learning environment	Feidakis et al., 2014
Behavioral Responses	
Facial expressions based on video	Postma-Nilsenová, Postma, & Tates, 2015
Facial expressions based on EMG	Durso et al., 2012
Overall body movements	Caballe et al., 2014
Body posture based on chair sensors	D'Mello & Graesser, 2012
Conversational cues	D'Mello, Craig, Witherspoon, McDaniel & Graesser, 2008
Eyetracking, changes in visual exploration	Graesser, Lu, Olde, Cooper-Pye & Whitten, 2005
Eyetracking, frequency and number of fixations	DeLucia, Preddy, Derby, Tharanathan, & Putrevu, 2014
Log-file triangulation in LMS	Pardos, Baker, San Pedro, Gowda, & Gowda, 2013
Physiological Responses	
EDA	Pecchinenda & Smith, 1996; Shen, Wang & Shen, 2009
fMRI	Hester, Barre, Murphy, Silk, & Mattingley, 2008
EEG	Wang et al., 2013
Pupillometry	Umemuro & Yamashita, 2003

As a last physiological response, the measurement of pupil dilation (pupillometry) is discussed, which most of today's eye trackers are capable of. Although one study could detect 75% of the confusion-induced trials in a problem-solving task (Umemuro & Yamashita, 2003), the method seems to be too sensitive, as other emotions (e.g., Bradley, Miccoli, Escrig

& Lang, 2008) and CL (e.g., Palinko, Kun, Shyrokov & Heeman, 2010) were found to be related to pupil dilation and thus, the method has low discriminatory power.

Drawbacks of Current Confusion Detection and Measurements

The wide range of used data streams and analytic approaches already shows that recognizing confusion is not straightforward. Evaluating and comparing the mentioned data sources and methods is difficult. It is hardly possible to compare performance of predictions or the amount of variance that has been explained by a certain data channel as the mentioned papers all use different labels for confusion, e.g., self-reports or a combined measure of self-reports and human-coded facial expressions. Not only does the measurement of confusion differ, but also how confusion is operationalized for the induction in experimental settings, and in what context it occurs when investigated in field settings. Hence, instead of comparing explained variance and model performance, the advantages and drawbacks of each can be described that build a trade-off for using them for different purposes. For example, fMRI can obviously not be used in the field but can contribute to fundamental theory building, while patterns in user interaction might not be generalizable to a theoretical framework level but are applicable in an actual learning environment in the field.

Probably the most important drawback for the purpose of supporting students in adaptive technology enhanced learning environment is that most of the mentioned data sources are only available inside a lab, as considerably expensive hardware is needed to acquire the data. However, even if the hardware was available, some data channels suffer from low sensitivity or specificity. Moreover, attaching instruments to learners is obtrusive and students may have privacy concerns. Therefore, measures might disturb the learning process and can be reactive to the variable of interest, as well as the learning outcomes. Regarding self-reports of emotions, measures typically suffer from interference with the learning task, as well as social biases (e.g., willingness or honesty to report confusion) and cognitive biases (e.g., not being aware about certain emotions). Another issue is that most measures focus on the detection of whether learners are confused or not (or the probability of being confused), and are not capable of reporting about the level of confusion - although the theoretical construct of confusion is a continuum rather than a binary state. Measuring frequency, duration and intensity of confusion would be important to predict whether the confusion will have a positive or negative impact on learning and decide for subsequent interventions (Arguel et al., 2017). Aiming at higher power, sensitivity, specificity, and time resolution, multimodal and multichannel approaches with models that take the data of several sources into account are discussed and examined as a solution for some of the mentioned issues (D'Mello & Kory, 2015; Hussain, AlZoubi, Calvo & D'Mello, 2011).

Detecting Confusion Based on Mouse Behavior

The issue of measurements being only available inside the laboratory can be addressed by using data on the observable interaction between the learner and the learning environment which is recordable without any additional hardware or special software (e.g., mouse and typing behavior). Hence, it is a potential unobtrusive and non-reactive data source that is available in the field, and in real-time. However, such data can only act as a proxy for any psychological latent variable. Thus, empirical studies need to investigate the relations between this data and latent variables (see chapter 4.1). Existing studies examined the relation between mouse behavior and boredom (Tsoulouhas et al., 2011), anxiety (Yamauchi, Seo, Choe, Bowman & Xiao, 2013), perceived need for help (Attig, Then & Krems, 2018), and valence and arousal (Maehr, 2005; Salmeron-Majadas, Santos & Boticario, 2014; Sottilare & Proctor, 2012; Zimmermann, Gomez, Danuser & Schär, 2006) with mixed results. Regarding confusion, there is no sufficient evidence yet on whether / how mouse behavior relates to it. Pentel (2015) aligned features of mouse movement data (e.g., directions, direction changes, changes in speed) to confusion states identified in think aloud protocols during a number game. Different classifiers for a player being confused or not reached accuracy rates of up to 94%. Although this seems very promising, there are some limitations regarding these results: 1) the artificial paradigm and context of this game inherently leads to a very specific pattern of mouse movements, and results are therefore not generalizable to navigation in web environments, 2) coding of confusion in the think aloud data was very broad (e.g., “I saw it before, but now it is not there anymore.” was coded as confusion), and 3) confusion and frustration were both treated as confusion. In another study, mouse features (left click rate, double click rate, tome to first left click, time to first double click), gaze and pupil features, and head position features were used in a random forest classifier and being confused was reported by users through clicking on a “I am confused” button (Lallé, Conati & Carenini, 2016). However, none of the included mouse features were in the top 10 features leading to a sensitivity of 61% and a specificity of 92%.

As shown in the mentioned studies, the operationalization, induction and manipulation of confusion for experimental research is not straightforward. In another recent study in the area of survey methodology, mouse behavior while filling a multi-item single-choice survey has been examined as a possible method to detect “whether a respondent is having trouble answering a question and what is causing their confusion” (Horwitz, Kieslich & Kreuter, 2017, p. 9). Although the “trouble” that respondents have during a survey was framed as confusion, the authors compared the mouse behavior for answer options using a straightforward wording against using complex wording. It is argued that added “repetitive, bureaucratic, technical information” (Horwitz, Kieslich et al., 2017, p. 12) in the complex version leads to confusion. However, complex answer options do not (necessarily) induce confusion in the theoretical understanding of an unresolved cognitive disequilibrium that was mentioned earlier. Instead, varying the complexity of the wording of answer options rather manipulates the experienced CL. A more complex wording of an answer option makes it more difficult to extract the relevant information, leading to an increased ECL. Although the operationalization is problematic as it does not cover confusion, and no effect sizes were reported (and could not be calculated due to missing test characteristics), the outcomes of this study are interesting: more and longer hovers were found for complex worded answer options then for the straightforward options.

In another study conducted by the same first author (Horwitz, Kreuter et al., 2017), the relation between mouse movement patterns while answering, and the reported item difficulty was examined. Participants were asked to answer a single-choice question based on a randomly assigned description of a scenario, which was manipulated to be formulated either complicated or straightforward. After answering, participants should self-report the difficulty (5-step Likert-scale) of the item. The study examined the relation between the reported difficulty and the mouse patterns while answering. Specific mouse movement patterns that were identified as frequently occurring in previous research were investigated. These are horizontal tracking, vertical tracking, hovering, using the mouse as a marker, and regressing between two areas of interest. Horizontal and vertical tracking refers to the mouse following the gaze position in the according direction. “Hovering” was defined as holding the mouse cursor over the question for more than 2 seconds, “marker” as holding the mouse cursor over an answer option for more than 2 seconds, and “regressive” described a move back and forth between two of the elements “question”, “answer option”, “white space”, and a “next question” button. Significant relations were found between three of the patterns (hover, marker and regressive) and the reported difficulty. Adding the three mouse movements patterns to a model led to a slightly higher predictive power (AOC = 2,119.74; ROC = 0.7911) compared a model that only includes response times for an item (AOC = 2,142.75; ROC = 0.7798). One advantage of the used approach that the authors did not mention is that response times are usually not available for items, if more than one item is presented on a page. Even if one item per page is presented, mouse behavior provides more valid information than simple logfiles on navigation (see chapter 4.1.1.3). The aim of the authors was to identify respondents that face difficulties while answering and therefore, "delivering help to confused respondents in real time and as a diagnostic tool to identify confusing questions" (Horwitz, Kreuter et al., 2017, p. 1). Again, experiencing difficulty while answering an item was equated to being confused. Another drawback that limits the use of the approach in the field is the use of manual coders that had to watch screen recordings to identify the mentioned mouse patterns. This was necessary as the used methods to record mouse movements did only cover pixel coordinates, but did not include to what element the positions refer. Thus, the recognized patterns of mouse movements could not be detected automatically in real-time. Moreover, only mouse movements were considered, but the selection of answers (represented as clicks) was ignored. Therefore, it is not possible to include some important indices such as the time to first answer selection. These drawbacks are addressed in this study by using the peripheral data approach (see chapter 4.1).

4.4.1 Research Questions and Hypotheses

Research Question 1: Detecting and Measuring Confusion

Research on the detection and measurement of confusion currently lacks a method that is applicable in online learning environments outside the lab. Promising attempts to use mouse movements either used ecologically invalid artificial environments (Pentel, 2015), failed at correctly operationalizing confusion based on existing theory, or neglected mouse behavior besides movements (Horwitz, Kieslich et al., 2017; Horwitz, Kreuter et al., 2017).

In the study of this work, the detection of confusion through mouse behavior during the interaction in tasks that are relevant for learning in advanced learning environments is investigated. Used in described studies (Horwitz, Kieslich et al., 2017; Horwitz, Kreuter et al., 2017), filling single-choice items provides a suitable task for this study for several reasons: 1) the structure, including a question and answer options, limits possible mouse behavior options, 2) answering single-choice items is an ecologically valid, relevant, and common task in online learning environments, e.g., used in rapid assessment tasks (e.g., Kalyuga, 2008; Renkl, Skuballa, Schwonke, Harr & Leber, 2015) or prompts (e.g., Bannert, 2007), and 3) the structure allows for a theory-based manipulation that induces confusion. If evidence is found for mouse behavior being a valid proxy for confusion within the simplified, limited structure of multi-item scales, this can lead to a direct application, but also, the approach can be examined in a less limited, less structured environment in order to generalize the results.

The first research question addressed in this study is therefore: Do indices of mouse behavior during the answering of single-choice items in questionnaires correspond to confusion?

The hypotheses for this research question are derived from the latter mentioned advantage of the structure of single-choice question that allow for a theory-based manipulation. D´Mello and Graesser (2014) describe that interruptions of organized sequences of actions induce confusion. In single-choice questions, the same sequence of actions is required for each item: Reading the question, deciding for one of the five presented answer options, and clicking on it. Hence, it is argued that confusion can be induced by interrupting this sequence of actions. Such an interruption that leads to confusion can be achieved through enriching items with an answer option that contains contradictory information (Arguel, Lockyer, Kennedy, Lodge & Pachman, 2018). This claim is further supported by Mandler's interruption (discrepancy) theory (Mandler, 1990). He argues that attention of individuals shifts towards discrepant information when detecting them during the assimilation of new information. This shift of attention should be reflected in how learners interact with the discrepant information (i.e., manipulated answer option by including confusion through contradictory information). Hence, it is hypothesized that this interaction, operationalized by different indices of the mouse behavior that is shown while answering single-choice questions (e.g., time spent on an answer option), will be higher for items with manipulated, confusing content (i.e., contradictory statements) compared to non-manipulated, non-confusing content (Hypothesis 1a: *Mouse-detects-confusion-Hypothesis*).

Discrepant information that triggers an interruption of a sequence of actions may not always lead to the same level of confusion or to confusion at all. The resulting cognitive or emotional state and its level of intensity might depend on the type of information that interrupts the sequence. In order to check whether the type of discrepancy is reflected in the mouse behavior, items were manipulated to be grammatically wrong. It is argued that grammar errors still interrupt a sequence and lead to confusion, but the intensity should be lower (which was checked in a pre-test of the items). Accordingly, it is hypothesized that indices of mouse behavior (e.g., time on item) for items with a grammatically wrong answer option are significantly lower than items including contradictions but are still higher than items without any manipulation (Hypothesis 1b: *Mouse-measures-confusion-levels-Hypothesis*).

Research Question 2: Relations Between Mouse Behavior and Item Difficulty

Moreover, this study seeks to confirm and extend the results of Horwitz and colleagues (Horwitz, Kreuter et al., 2017), that mouse behavior is related to the subjective difficulty-ratings of knowledge items. Although the aim of the mentioned study was to identify respondents with trouble in online surveys, their research question is highly relevant for the area of learning in online environments. Rating the difficulty of an item requires metacognitive activity and represents a metacognitive task. A learner needs to reflect about his own knowledge on the question and evaluate how this relates to the population or a specific sample. Hence, if it is possible to replace the self-report of the perceived difficulty of an item through indices of mouse behavior while answering the item, the metacognitive activity becomes measurable at least to some degree. Although this is only one aspect of metacognition in a special context with limited generalizability, finding a measure for it is very valuable, taking into account the fundamental issues that researches have continuously been reporting for decades regarding the measurement of metacognition (e.g., Veenman et al., 2006). In online learning environments, being able to assess the difficulty that learners have with answering a single-choice question would add considerable insight, compared to only knowing whether the given answer was right or wrong. As an example, a derived instructional possibility could be an adaptive restudy that not only considers wrong questions but also questions that were right but still perceived as difficult, according to the interaction with it.

In addition to this, this study is also interested in the relation of mouse behavior to the objective difficulty of items. As learners tend to make wrong judgments towards an overestimation of their performance and abilities (e.g., Dunlosky & Lipko, 2007), validating these judgements with their interaction could contribute to higher accuracy.

Summing this up, the second research question of this study is: Is the perceived and the objective difficulty related to mouse behavior while answering single-choice items of general knowledge? From this question, four hypotheses are derived

Firstly, it is hypothesized that indices of mouse behavior positively correlate with the level of reported subjective difficulty (Hypothesis 2a: *Higher-Mouse-Higher-Subjective-Difficulty*). Moreover, it is hypothesized that these mouse indices can predict the subjective difficulty in a regression model (Hypothesis 2b: *Mouse-indicates-subjective-difficulty-Hypothesis).*

Regarding objective difficulty, it is hypothesized that indices of mouse behavior also positively correlate with it (Hypothesis 2c: *Higher-Mouse-Higher-Objective-Difficulty-Hypothesis*), and that indices of mouse behavior predict the correctness of an answer in a binary logistic regression model (Hypothesis 2d: *Mouse-indicates-objective-difficulty-Hypothesis*).

Research Question 3: Correspondence of Mouse Behavior with Feeling-of-Knowing Judgements

In addition to the perceived and objective difficulty of an item, this study addresses the feeling-of-knowing (FOK) for the items as a metacognitive judgement. Such judgments describe the predictions made by an individual to be able to recall a given, specific information from their existing knowledge ("I know the answer of this question" versus "I do not know the answer of this question"). As such, FOK judgements do not refer to an actual answer to a question. It is argued that FOK judgements can be helpful in adaptive learning environments to check whether learners already understand an entity of the curriculum. Glucksberg and McCloskey (1981) found that don't-know responses can be made quickly and accurately when no relevant information is known. In contrast, a don't-know response is slow if some relevant knowledge is available, because the person needs time to evaluate if he/she has enough knowledge or is sure enough to state that he/she knows the answer. As the selected questions in the BEFKI are general knowledge questions, it is argued that the participants should have some prior knowledge of these questions. However, very difficult items were added of which participants probably do not have prior knowledge. Thus, the third research question in this study is: Are response times of FOK judgements related to the subjective difficulty of single-choice items of general knowledge? For this question, it is hypothesized that response times for FOK judgments are positively related to the subjective difficulty rating (Hypothesis 3a: *Slower-FOK-Higher-Difficulty*). Moreover, it is hypothesized that for questions of very high subjective difficulty, the relation is inverted (Hypothesis 3b: *Faster-FOK-for-Extreme-Difficulty*).

4.4.2 Method

4.4.2.1 Sample and Design

A correlational online field study was conducted with $N = 144$ university students (46% male, age $M = 23.26$, $SD = 2.45$). For participant recruiting, advertisement was posted in three Facebook groups of different German universities. The advertisement involved the following information: participation takes approximately 20 to 30 minutes, the task is to “answer questionnaires about your personality and your general knowledge”, requirements to participate are a calm environment, a laptop or desktop computer (no tablets or mobile phones), and 5 Euro will be payed or donated to a charitable as a compensation for a complete participation. It was checked for every participants whether they 1) spent a minimum of 15 minutes on the study, did not show non-meaningful answer behavior (e.g., always selecting the answer option of the same position), and completed all parts of the study. After removing participants that did not fulfil these requirements, 114 participants remained in the data analysis.

4.4.2.2 Research Paradigm

Although the final goal of this area of research is the detection of confusion during learning, independent of the structure and context of the learning environment, this study uses multi-item scales instead of a traditional learning environment. The reason for this choice were already mentioned in detail above: the structure limits the scope of possible mouse behavior, while still being a relevant task in online learning environments, and the structure allows for a theory-based manipulation that induces confusion. Limiting the structure of possible mouse behavior is an important part of the research paradigm in this study. Mouse behavior on websites directly depends on its design. For example, when content sections of a text are presented in different tabs, learners have to click to the corresponding tab in order to read it.

In comparison, splitting all content sections of the text by including sub headings, learners do not have to click. This leads to two fundamentally different patterns of mouse behavior that are represented in the recorded data and therefore, allows equally different inferences about learner's behavior and experience. In this example, opening a tab and spending enough time at it could be operationalized as reading the paragraph. Using this rationale, researchers can control the granularity and meaningfulness of recorded mouse behavior regarding a variable of interest by well-aimed decisions on how the information is presented. Therefore, how information in a learning environment is presented becomes an important decision during the planning of the research design of studies. A similar rationale has been introduced years ago as a "poor man's eye tracker" to get coarse information on what learners were reading (Ullrich & Melis, 2002). Manipulating the design to examine a specific question limits the generalizability of the results to the required interactions. However, this drawback comes with an important advantage: using the method allows to empirically discover and proof fundamental theoretical relations in a clearly defined and controlled scope.

The research paradigm of this study uses single-choice items. Single-choice items already limit the possible mouse interactions by their very nature. A minimum interaction is required to answer it: scroll to the question, move the mouse cursor over the answer option, click on the answer option. It is argued that everything beside this required interaction can potentially tell about latent variables. Relations between the interaction and that latent variable can then be uncovered in two ways. First, items can be manipulated to change a latent variable. If the interaction with that item significantly differs compared to the mean interaction with non-manipulated items, then there is a relation that can be further investigated. For example, in this study, items were manipulated to induce confusion through contradictions (e.g., "I'm a tidy person, [not] cleaning up often"). Secondly, relations between the interaction and meta-information of items can be checked. This meta-information can either be inherently available (e.g., the objective difficulty of an item defined by how many subjects were able to solve it), or needs to be acquired (e.g., ratings of subjective difficulty or FOK). It is important to understand, that the actually measured variables using the questionnaires (in this study, Big Five personality traits or crystalline intelligence) are not relevant to the questions of this study, but that the process data on how it was answered is. The questionnaires were selected because they are readily available, frequently used, and validation studies with measures of quality criteria such as reliability or selectivity are available.

4.4.2.3 Indices of Mouse Behavior

Combining the described research paradigm with the peripheral data approach allows the automated extraction of indices of mouse behavior. The developed software framework ScreenAlytics (see chapter 4.1), was used to record fine-grained data on the mouse behavior. It allows us to easily extract important indices of the mouse behavior from the interaction data. As in eye tracking methodology, different elements of the web-based questionnaires define areas of interest (AOI) as shown in Figure 19. From these AOI, indices of mouse behavior can be derived.

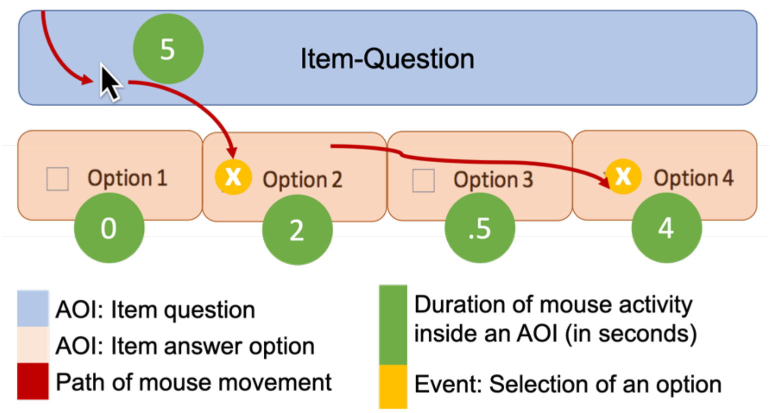


Figure 19. Elements on the website represent different areas of interest from which indices of mouse behavior can be extracted.

Used indices that can be drawn from the interaction with single-choice items are listed in Table 11. Transitions between elements can be counted automatically with information on the mouse movement paths. Compared to the indices used by Horwitz and colleagues (2017), their "hovering" refers to T-Q, "marker" to T-A and "regressive" refers to F-TWA, F-TAA and F-TAQ.

Table 11

Indices of mouse behavior regarding the interaction with single-choice items.

Unit	Code
Time …	
… on item question and answer	T-QA
… on item question text	T-QT
… on item answers	T-A
… till first selection	T-FS
Frequency of …	
… answer selections per item.	F-A
… transitions between white space and answers.	F-TWA
… transitions between different answer options.	F-TAA
… transitions between answer and question.	F-TAQ

Note. When referred to this table, codes are used e.g., T-QA for "Time on item question and answer".

4.4.2.4 Confusion Induction and Manipulation Check

Confusion was induced by manipulating items of the well-established German translation of the Big Five Inventory-2 questionnaire (see chapter 4.4.2.5 for a description of the questionnaire). The questionnaire consists of 60 5-step Likert-scaled items, divided into 6 pages with 10 items each. Within a page of 10 items, one item was manipulated to be either 1) contradictory or 2) grammatically wrong. It's position was randomly chosen.

In many studies, confusion (and its successful induction) is measured by self-reports during the learning process and thus, may be reactive to the variable of interest. In this study, items were pre-tested on whether and to what extent they induce confusion. By this, the induction of confusion is proofed but this manipulation check is not done during the assessment to prevent reactiveness. Eight participants were asked to rate the items on a 4-step Likert-scale from not confusing (0) to very confusing (3). Wilcoxon signed rank tests were conducted for each manipulated item to check whether the according confusion rating is significantly higher than those of non-manipulated items. Results are reported in Table 12.

Table 12

Description and check of the manipulated items for inducing confusion.

Type of manipulation	Original Version (item position in questionnaire)	Manipulated Version	Pre-test confusion rating (N=8)	
			M (SD), *p*	*z*, Cohen's d*
Contradiction	I stay relaxed even in stressful situations. (4)	I stay calm even in relaxed situations.	2.63 (1.06), $p < .05$	-2.380, 3.115
Grammar error	I'm systematic, keeping my things in order. (18)	I systematic am, keeping my things in order.	2.63 (0.52), $p < .05$	-2.521, 3.932
Contradiction	I am confident, satisfied with myself. (24)	I'm confident, dissatisfied with myself.	2.88 (0.35), $p < .05$	-2.521, 3.932
Grammar error	I'm efficient, I do things fast. (38)	I'm efficient, does things fast.	2.37 (1.06), $p < .05$	-2.380, 3.115
Contradiction	I'm more of a mess. I rarely clean up. (48)	I am rather neat. I seldom clean up.	2.63 (0.74), $p < .05$	-2.383, 3.128
Grammar error	Sometimes I act irresponsible, reckless. (58)	Sometimes act irresponsible, reckless.	1.25 (1.06), $p = .313$	-1.120, .862

Note. Items were translated from German and contradictions and grammar error might therefore not represent the same quality in English. * This value represents a z-transformation of Wilcoxon's W-value.

Although grammar manipulation of the last item was not significant, it has a high effect size (according to Cohen, 1988) and was still kept due to the low statistical power that the pre-test had. Pre-testing the items also showed that grammatically wrong items tend to induce less confusion ($M = 2.54$, $SD = .47$), than contradictory items ($M = 2.70$, $SD = .37$) with a medium effect size ($d = .723$), although not significant in a Wilcoxon test, again due to the low power of the pre-test (N = 8, $z = -.962$, $p = .336$). This difference means that different levels of confusion can be induced by the used manipulations which is a requirement to identify them through mouse behavior (*Mouse-detects-confusion-Hypothesis*).

In addition to this check, participants were asked to provide suggestions on how to improve the design of the study after completion, but before the debriefing. The question did not involve any cue to focus on a specific part of the study. These open answers were analyzed for mentioning anything related to the manipulated items. Answers of 118 participants that answered the open question were analyzed. Of these 118 participants, 15 participants only mentioned grammar errors, 52 only mentioned contradictions, and 22 mentioned both. Seventeen participants answered that there is nothing to improve. Hence, 75% of the answers mentioned contradictions whereas only 31% mentioned grammar errors. This further supports the assumption that grammar levels induced less confusion than contradictory items.

4.4.2.5 Measures and Instruments

Adapted BEFKI GC-K

BEFKI GC-K is a short, 12-item knowledge scale to measure crystalline intelligence (g_c) using declarative knowledge items from the sciences, the humanities, and civics (e.g., “What symptoms are typical for epilepsy?” or “What does amber consist of?”). It is based on the item pool of the “Berliner Test zur Erfassung Fluider und Kristalliner Intelligenz” (BEFKI, berlin test to assess fluid and crystalline intelligence) project, which has been validated on a representative Sample of 1134 German adults (Schipolowski et al., 2013, 2014). To the existing 12 questions, a total of six items were added to cover very low and very high difficulty, as listed in Table 13. Reported reliability of the validation study is Cronbach’s Alpha = .70 to .82. In this study, the items reached a reliability of *Cronbach’s Alpha* = .57 including the extreme items and *Cronbach’s Alpha* = .54 when only taking the original items into account.

Table 13

Items with High and Low Difficulty Added to BEFKI

#	Difficulty	Item
1	Low	How many federal states has Germany?
2	Low	What is the name of a famous comic elephant with big ears?
3	Low	Which actor later became US president?
4	High	Who invented the microphone in 1878?
5	High	When did the broadcast of color television in Germany begin?-
6	High	In which town was Marilyn Monroe born?

This scale was used in two variants: first, only questions without answer options were presented together with a binary FOK judgement (“I know the answer” and “I don’t know the answer” options). Moreover, an additional 5-step Likert scaled judgement of the perceived subjective difficulty was presented for each item (“I think the above question is very easy / rather easy / medium / rather difficult / very difficult”). Figure 20 shows the design of the measurement of the subjective difficulty rating and the feeling-of-knowing judgement.

Secondly, at a later point in the data acquisition, the questions were presented again as single-choice items with 4 answer options, one being the correct answer to measure the objective difficulty of the item over all participants. Figure 21 shows the design of an example question with 4-answer options.

Objective difficulty for the items was calculated as the ratio between participants that got the item right and the total number of participants and is shown in Table 14, ordered by their difficulty. As expected, the added items intended to have “high” difficulty, was answered correctly by the smallest proportions (23, 35 and 46,2%). Regarding the added items that were intended to have “low” difficulty, only two of the items actually had the expected low difficulty (94,9%). The item “Which actor later became US president?” had a rather high difficulty of 76,1%.

Table 14

Items of the general knowledge scale ordered by difficulty.

Position in questionnaire	Question	Participants that answered correctly (of N=114 in total)	
		Frequency	Percentage
18	In which town was Marilyn Monroe born?	27	23
16	Who invented the microphone in 1878?	41	35
17	When did the broadcast of color television in Germany begin?	54	46.2
15	What happened after the "Battle of Leipzig"?	64	54.7
14	Family and inheritance law is subject of what?	73	62.3
11	What is the characteristic of a diode?	76	65
12	What's the "Nibelungenlied"?	78	66.6
13	What are royalties?	80	68.4
1	Which actor later became US president?	89	76.1
10	What is nihilism?	89	76.1
7	A well-known painting by Dalí shows "melting Clocks". Which style can be assigned to this painting?	91	77.7
5	What was the task of the Inquisition courts of the Middle Ages?	93	79.5
8	What's mitosis for?	95	81.2
9	What's a petition?	104	88.9
6	What is amber made of?	108	92.4
4	Which symptoms are typical for epilepsy?	110	94
2	How many federal states does Germany have?	111	94.9
3	What is the name of a comic elephant with large ears?	111	94.9

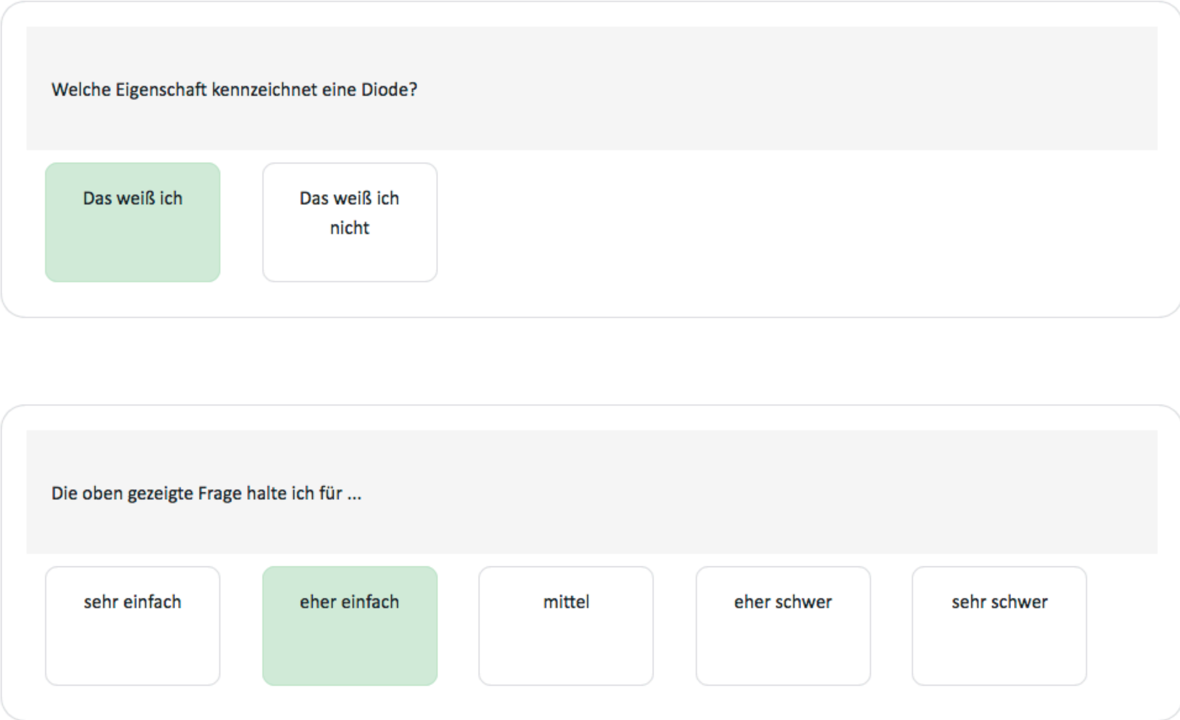


Figure 20. Measurement of binary feeling-of-knowledge judgement (question "What characterizes a diode?" with answer options "I know that" vs. "I don't know that"), and perceived subjective difficulty in a 5-step Likert-scale.

Figure 21. Measurement of objective difficulty by checking for the actual knowledge with a 4-option single choice item.

Big Five Inventory 2 (BFI-2) for Confusion Induction

BFI-2 (Danner et al., 2016) is a German version of the 60-item Big Five Inventory 2 that measures the big five personality traits extraversion, openness to experience, conscientiousness, agreeableness and neuroticism. Reported reliability of the sub-scales in the

German validation study is Cronbach's Alpha = .70 to .80. In the study of this work, the items were divided into 6 pages, each presenting 10 items. The order of the items was kept as in the original version. On every page, one of the 10 presented items were manipulated to induce confusion by either making them contradictory (e.g., "I am tidy, *not* cleaning up often) or including grammar errors (e.g., wrong verb position or wrong cases). The detailed manipulation is described in chapter 4.4.2.4.

4.4.2.6 Procedure

The procedure of this study is listed in Table 15. On the initial webpage, participants were briefed about 1) the requirements to receive the compensation payment, 2) interaction data being collected during the study, time that the study will approximately take (20 to 30 minutes), technical requirements (using a desktop device, no reloading or leaving of the website, no use of browser navigation buttons, maximized browser window), experimental requirements (e.g., no parallel interaction with Facebook or Google) and anonymization of the acquired data. On the same page, participants were asked for demographics (age, sex, occupation) and their confirmation of the following statements: 1) I have 30 minutes time now to participate in this study without breaks, 2) I work on a desktop computer or a laptop and my browser is maximized, 3) I did not yet participate in this study, 4) I agree that my anonymized answers are stored for data analysis, 5) I am over 18 years old and I have a bank account for the transfer of my compensation payment of 5 Euro. After that, all FOK and subjective difficulty ratings for the 18 items of the adapted BEFKI scale were presented on one page. Participants then filled the 60 items of the manipulated BFI2 scale divided on 6 page, 10 items each. Then, participants were asked to give the actual answer to the single-choice items of the adapted BEFKI scale, which was rated before regarding FOK and difficulty. An open answer text form was then presented and participants were asked to fill in suggestions to improve the study design. This was used to check whether participants recognized the manipulation of the BEFKI items. On the last two pages, bank information for the payment of the compensation was acquired and participants were debriefed about the study and the correct answers for the BEFKI items.

Table 15

Procedure and Instruments with Manipulations of Study 3

Page	Description	Variable(s)	Instrument	Manipulation	Hypothesis
1	Briefing, consent and demographics	age, sex, occupation, consent	-	-	-
2	Prior judgements/rating of general knowledge questions	FOK, difficulty ratings	FOK and difficulty rating of adapted BEFKI	-	2a, b / 3
3	Confusion induction and measurement	BIG5, mouse behavior	BFI-2, Item 1-10	Item 4: Contradiction	1a, b
4	Confusion induction and measurement	BIG5, mouse behavior	BFI-2, Item 11-20	Item 18: Grammar error	1a, b
5	Confusion induction and measurement	BIG5, mouse behavior	BFI-2, Item 21-30	Item 24: Contradiction	1a, b
6	Confusion induction and measurement	BIG5, mouse behavior	BFI-2, Item 31-40	Item 38: Grammar error	1a, b
7	Confusion induction and measurement	BIG5, mouse behavior	BFI-2, Item 41-50	Item 48: Contradiction	1a, b
8	Confusion induction and measurement	BIG5, mouse behavior	BFI-2, Item 51-60	Item 58: Grammar	1a, b
9	Answers to general knowledge questions	Objective difficulty, mouse behavior	Adapted BEFKI	-	2a, b / 3
10	Check feedback if BFI-2 manipulations were recognized	manipulation check	Open answer, suggestions and feedback	-	-
11	Compensation money	bank information	-	-	-
12	Debrief and answers to BEFKI	-	-	-	-

4.4.3 Results

If not mentioned, type I error rate was set to .05 for analyses. IBM Statistics 25, PHP, Python, Microsoft Excel and R were used to extract, filter, aggregate, and analyze the data set.

Of all time-related mouse behavior indices, values over 45s were regarded as not related to answering the questions and thus, were removed. All mouse behavior indices of all items were checked for collinearity, which was defined as a correlation of Pearson's $r > .7$. However, the highest correlation between two indices was $r = .57$.

Preliminary Assumption

Testing the indices of mouse behavior for statistical significant differences between manipulated and non-manipulated items is problematic for several reasons: 1) Corresponding tests (e.g., repeated measures t-test or analysis of variance) become significant even for small differences because of the high number of cases and the high variance, 2) Tests have to be calculated individually for each index, which is on the one hand very time-consuming and on the other hand leads to a possible underestimation of the effect, 3) Increased mean values of the indices can be high due to deviating interactions of some participant that remain on an item for a very long time because of other reasons (e.g., distraction). At the same time the removal of outliers according to a fixed criterion (e.g., mean value +/- 2*standard deviation) or the comparison of the medians is not reasonable, as for manipulated items, more values at the right of the median are expected. Therefore, it is reasonable to additionally determine which indices of mouse behavior were higher than the median of non-manipulated items for each subject and each item. For this purpose, binary variables were computed for all indices of all items, indicating whether the value is above the median of the corresponding index for non-manipulated items (0 = below the median, 1 = above the median). These binary variables are then summed up into a conglomerate for each item (called K below).

$$
\begin{aligned}
K_{Item\text{-}N} \quad = \quad & +1 \mid \text{if} \; (\; T\text{-}QA_N \; > Median_{T\text{-}QA} \;) \\
& +1 \mid \text{if} \; (\; T\text{-}QT_N \; > Median_{T\text{-}QT} \;) \\
& +1 \mid \text{if} \; (\; T\text{-}A_N \; > Median_{T\text{-}A} \;) \\
& +1 \mid \text{if} \; (\; T\text{-}FS_N \; > Median_{T\text{-}FS} \;)
\end{aligned}
$$

This results in a single conglomerate per item that contains all indices of mouse behavior. It also solves the problem of high standard deviation by single, extreme values.

4.4.3.1 Hypothesis 1a: Mouse-detects-confusion-Hypothesis

In order to check whether manipulated items that induced confusion have higher indices of mouse behavior and hence, can be recognized by the mouse behavior of the participants, the listed mouse indices of manipulated items were compared to those of non-manipulated items. Sixty BEFKI items were presented on 6 pages, 10 on each page. As one of 10 items on each page was manipulated, it was checked whether mouse behavior indices were higher for this item compared to the 9 non-manipulated items. Table 16 shows which indices were higher for manipulated items than for all non-manipulated items on the same page. Regarding the time-related indices, except for T-QT on the first manipulated item 4 and for T-QA, T-QT, and T-A on the grammar-manipulated item 58, all indices were higher for the manipulated items then for all other, non-manipulated items on the page (i.e., manipulated items had the highest values for time-related indices). Frequency-related indices were only rarely higher for manipulated items compared to non-manipulated items. As an example, Figure 22 shows the mean T-FS for all items on page 4 but the first, which is missing as time to first select needs a previous item to be computed. Figure 23 shows the T-QA index for all manipulated items compared to the according non-manipulated items on the page.

Table 16

Indication of higher indices for manipulated compared to non-manipulated items.

Item	Manipulation	T-QA	T-QT	T-A	T-FS	F-A	F-TWA	F-TAA	F-TAQ
4	Contradiction	X		X	X				
18	Grammar	X	X	X	X		X		
24	Contradiction	X	X	X	X				X
38	Grammar	X	X	X	X				
48	Contradiction	X	X	X	X		X		
58	Grammar				X				

Note. X = Higher value of the manipulated item than for every other item on the same page, being significantly higher compared to the overall mean of all non-manipulated items. T-QA = Time on question answer, T-QT = time on question text, T-A = time on answers, T-FS = time till first select, F-A = Answer selections per item, F-TWA = transitions between white space and answer, F-TAA = transitions between answer options, F-TAQ = transitions between answers and question

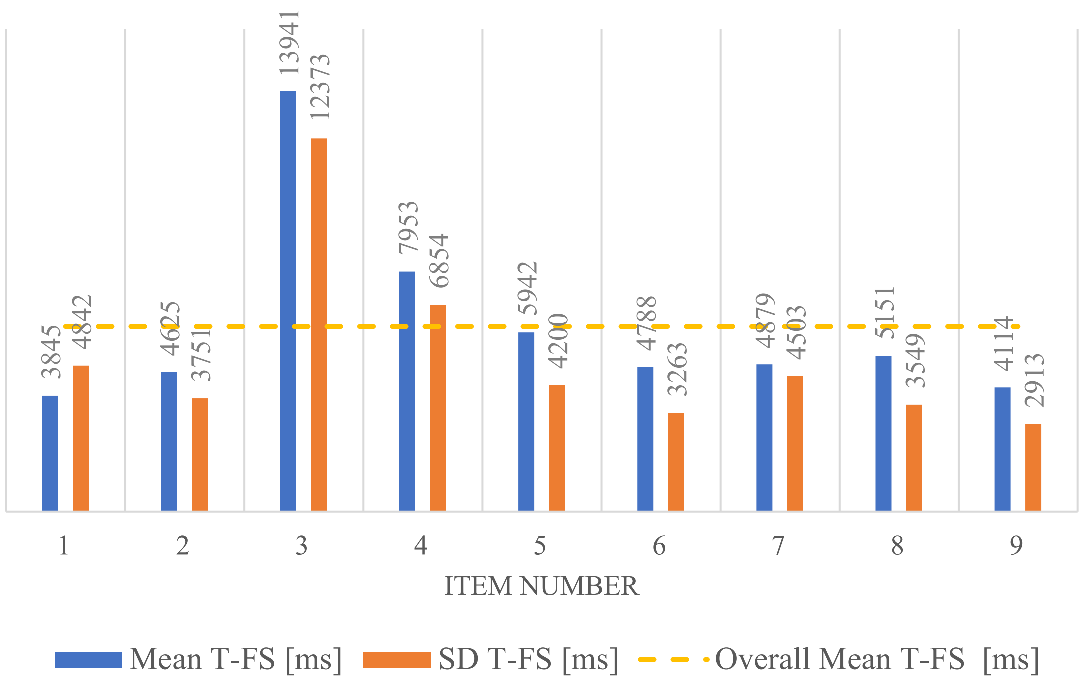


Figure 22. Time till first selection for 9 items of page 4. Item 3 was manipulated to induce confusion with a grammar error.

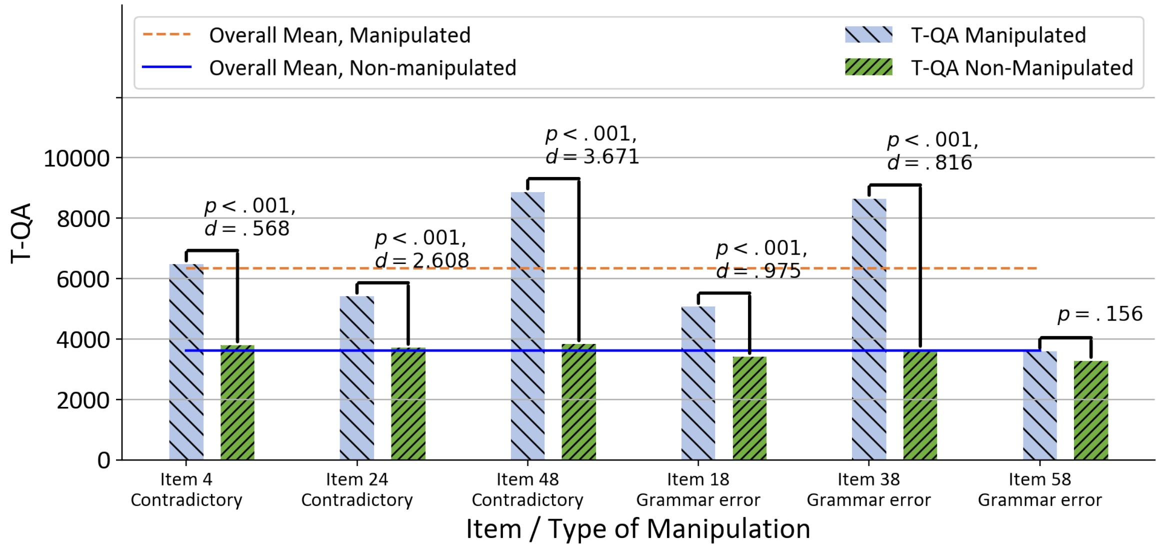
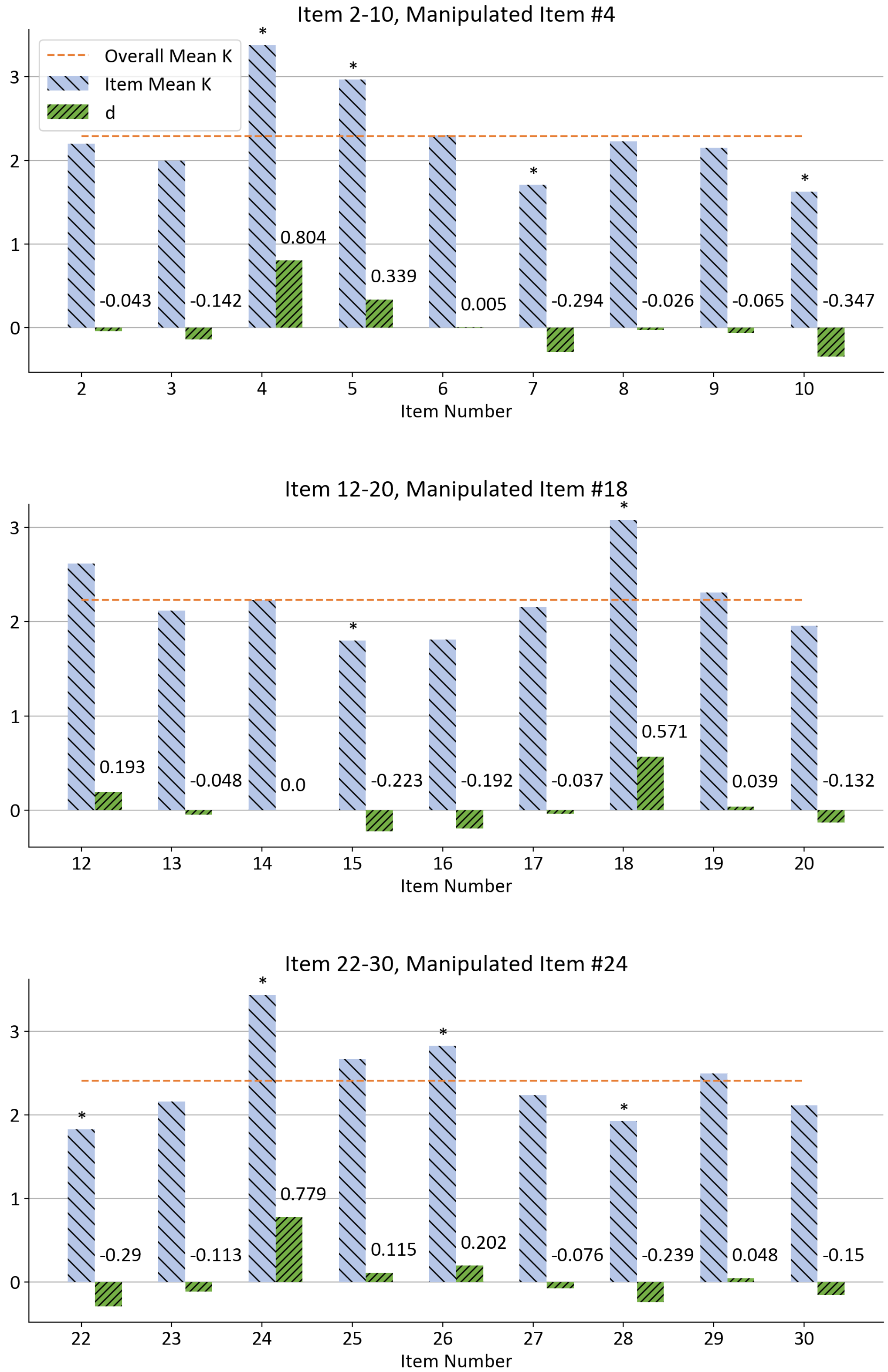


Figure 23. Mean T-QA of non-manipulated items vs. manipulated item on all six pages with 10 items each.

Although these results already tend to confirm the hypothesis, it lacks tests for statistical significance. As mentioned in 4.4.3, testing the mean value differences for statistical significance is problematic, and the computed conglomerate of mouse behavior indices K was used for significance test. Regarding the hypothesis check, it is expected that K is significantly higher for manipulated items than for non-manipulated items. The mean K-value over all items that were presented on one page was calculated and compared to the K of the individual items using paired t-tests. Cohen's d was calculated as effect sizes with a correction for paired t-tests as suggested by Morris (2008). Figure 24 shows six graphs including all items expect the first of every page as for the calculation of the conglomerate, a previous item was needed. All manipulated items could be identified by 1) being bigger than the overall mean K-value and 2) being significant, indicated with a star. Thus, the hypothesis can be confirmed.


Item 2-10, Manipulated Item #4
Overall Mean K
Item Mean K
d
-0.043
-0.142
0.804
0.339
0.005
-0.294
-0.026
-0.065
-0.347
Item Number
Item 12-20, Manipulated Item #18
0.193
-0.048
0.0
-0.223
-0.192
-0.037
0.571
0.039
-0.132
Item Number
Item 22-30, Manipulated Item #24
-0.29
-0.113
0.779
0.115
0.202
-0.076
-0.239
0.048
-0.15
Item Number

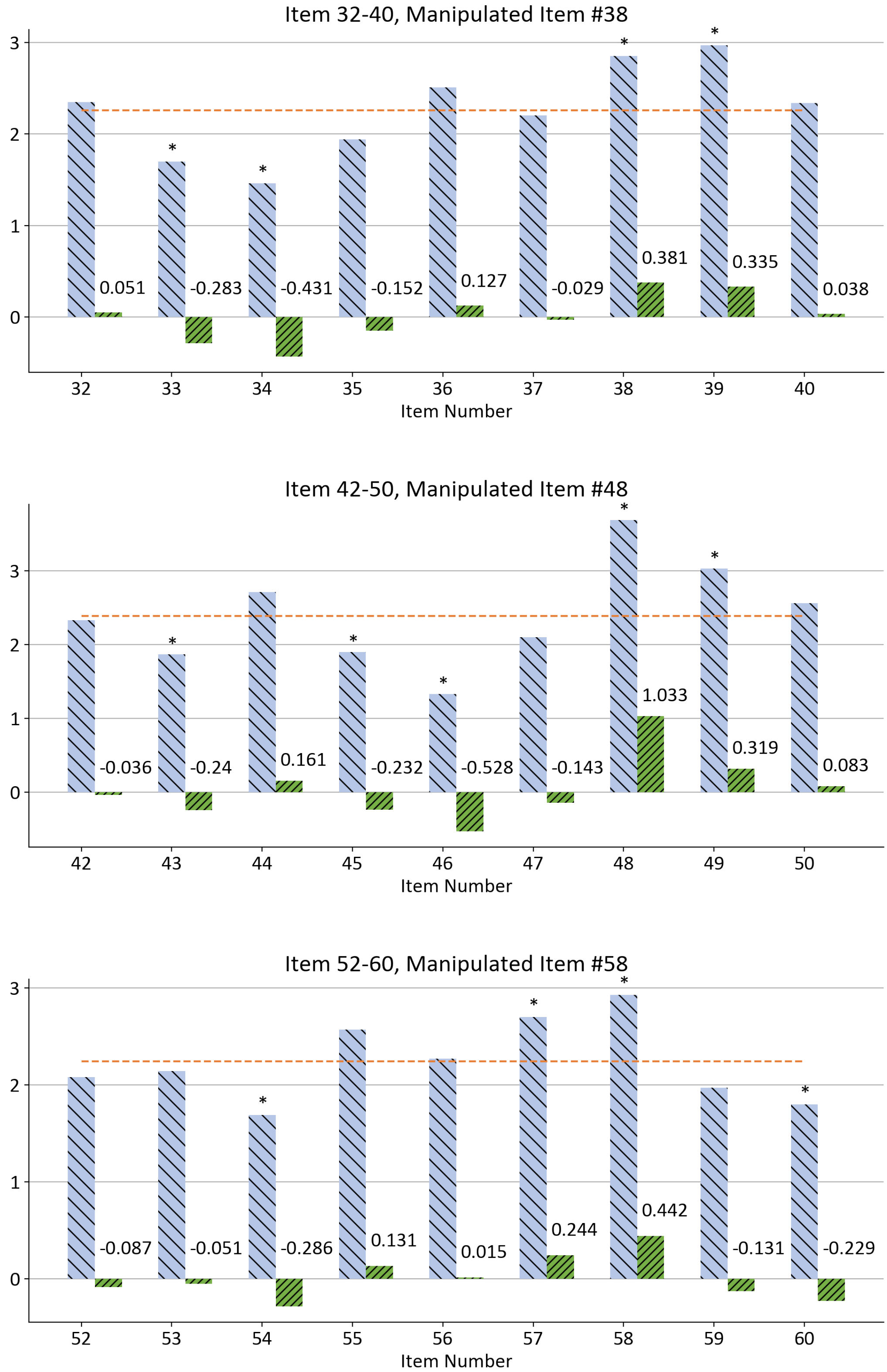


Figure 24. Comparison of manipulated and non-manipulated items regarding a conglomerate K of their received mouse behavior, grouped as 10 items were presented on 6 different pages. First items are missing as calculation of K requires a previous item. * = significant difference between K and overall mean K (p < .001); *d* = Corrected *Cohen's d* for paired tests. *N* = 115.

4.4.3.2 Hypothesis 1b: Mouse-measures-confusion-Levels-Hypothesis

It was argued that grammar errors induce less confusion than contradictions. Hence, it was checked whether items with grammar manipulation have significantly lower indices of mouse behavior than items with contradiction manipulations, but higher indices than non-manipulated items. The same conglomerate K of all time-related indices of mouse behavior as in the *Mouse-detects-confusion-Hypothesis* was used to compare the items. Mean K-values were built for manipulated items with contradictions, manipulated items with grammar errors and non-manipulated items. These values were compared by using paired t-tests. The result is shown in Figure 25. Contradictory items (M = 3.50; SD = 0.80) show significantly higher K-values than items with grammar errors (M = 2.95; SD = 0.89), t (114) = 5.704, p < .001 with an effect size of *Cohen's d* = .568. Moreover, items with grammar errors show significantly higher K-values than non-manipulated items (M = 2.30; SD = 0.16), t (114) = 8.186, p < .001 with an effect size of *Cohen's d* = .904. Thus, the hypothesis can be confirmed.

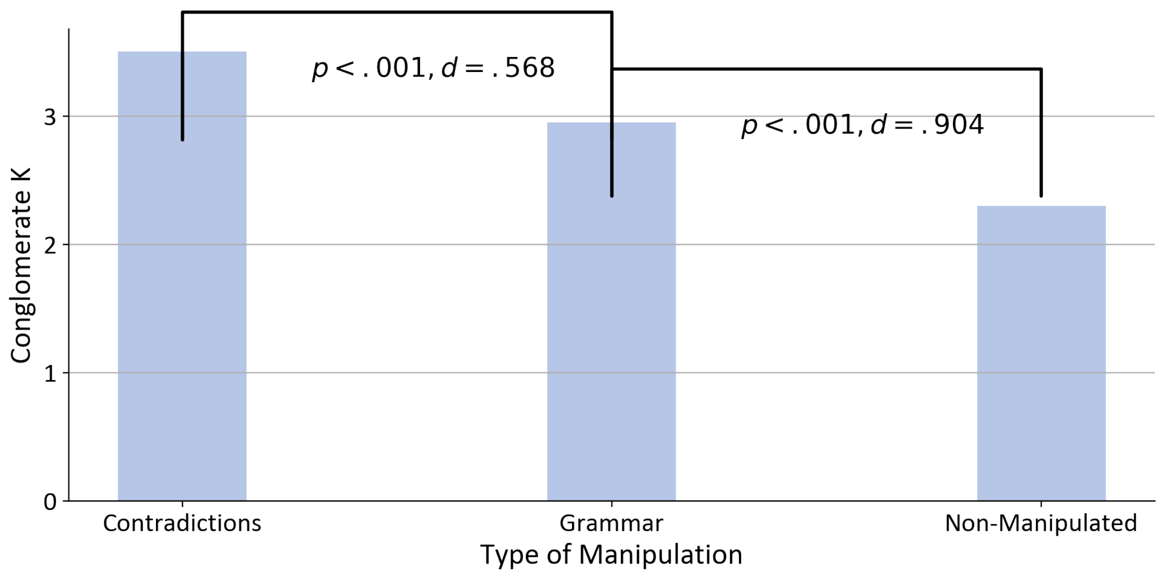


Figure 25. Comparison of manipulated items with contradictions, grammar errors and no manipulation regarding a conglomerate K of their received mouse behavior.

4.4.3.3 Hypothesis 2a: Higher-Mouse-Higher-Subjective-Difficulty

Mouse indices for each item were correlated with the 5-step Likert-scaled difficulty rating of the according item. It is important to understand that the mouse behavior during the actual answering of the item was assessed, not the mouse behavior during the rating of the item. For each mouse index, the number of significant correlations with the difficulty rating of an item was counted and a mean value was calculated as shown in Table 17. The number of positive correlations (17) is higher than the number of negative correlations (5), indicating that higher values of subjective difficulty relate to higher indices of mouse behavior (e.g., the higher the total time on the item, the higher its subjective difficulty). Moreover, correlations were computed for aggregated indices over all items, ignoring the item level listed in Table 18. Correlations are significant, with low correlations ranging between r = .061 and .120.

Table 17

Correlations between indices of mouse behavior on items and subjective difficulty rating.

	T-QA	T-QT	T-A	T-FS	F-A	F-TWA	F-TAA	F-TAQ
Number of items with sign. positive correlation	5	0	2	1	0	0	5	2
Mean r	.194	-	.270	.204	-	-	.181	.196
Number of items with sign. negative correlation	1	1	0	1	0	1	0	1
Mean r	-.173	-.169	-	-.174	-	-.167	-	-.167

Table 18

Correlations between aggregated indices over all items and subjective difficulty rating.

Index	T-QA	T-TQ	T-A	T-FS	F-A	F-TAA / F-TQA
r	.120**	.061**	.098**	.073**	.076**	.114**
n	2000	2003	1997	1731	2034	2034

Note. n for index T-FS is lower as the computation of the time to first selection needs a previous item and hence, could not be computed for the first item on a page. ** = significant on the 0.01-level.

In addition to this, indices of mouse behavior were aggregated over all BEFKI items and compared between the answer options of the subjective difficulty rating. For an easier comparison, all indices were mapped on values between 1 (minimum) and 10 (maximum). As shown in Figure 26, higher values of the indices T-A and a combination of F-TAA/F-TQA (transitions related to an answer option) correspond to higher subjective difficulty ratings. When ignoring the "very difficult" rating, this is true for all indices but F-A. Conducted ANOVAs revealed significant difference between the rating levels for all indices, as shown in Table 19. Considering the significant correlations between items' subjectivity ratings and indices of mouse behavior, and the significant correlations for all indices when ignoring the item level, this *Higher-Mouse-Higher-Subjective-Difficulty* hypothesis can be accepted.

4.4.3.1 Hypothesis 2b: Mouse-indicates-subjective-difficulty-Hypothesis

To check whether the mouse behavior indicates subjective difficulty of general knowledge items in the BEFKI questionnaire, multiple regression models were computed for items with at least two significant indices entering the significantly correlated indices of mouse behavior as predictors and subjective difficulty as the dependent variable. As Table 20 shows, more than one index was significantly related to the subjective difficulty of the items 5, 11 and 12 with an explained variance between 2.5% and 9.6%.

Due to the low explained variance by the indices in the multiple regression model, an indication of subjective difficulty by indices of mouse behavior does not seems to be reliable. Therefore, the hypothesis is rejected.

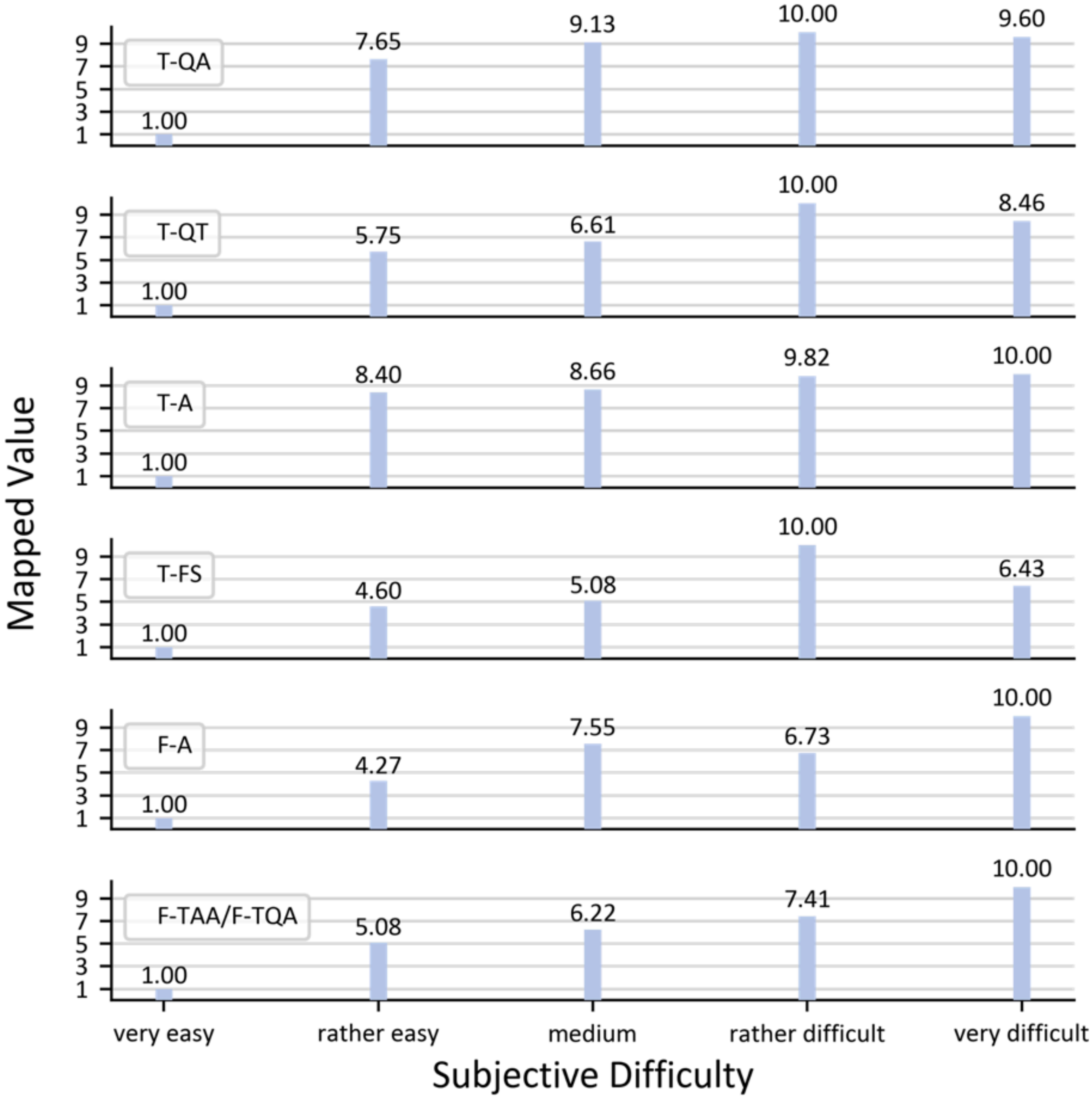


Figure 26. Indices of mouse behavior, aggregated over all BEFKI items and compared between answer options of subjective difficulty rating. For standardized comparison, values were mapped on a scale from 1 (minimum) to 10 (maximum).

Table 19

Results for ANOVAs checking the levels of subjective difficulty rating for significant differences on indices of mouse behavior.

Index	T-QA	T-QT	T-A	T-FS	F-A	F-TAA/F-TQA
F	10.954	2.398	7.440	3.561	3.581	7.149
df	4, 1995	4, 1998	4, 2029	4, 1728	4, 2029	4, 2029
p	< .001	< .05	< .001	< .01	< .01	< .001

Table 20

Regression models for subjective difficulty of items with indices of mouse behavior as predictors.

Subjective Difficulty of Item	Model / Predictors	sig.	df1, df2	R^2_{adj}	F
5	T-A, T-FS	.047	2, 106	.038	3.144
11	T-QA, F-TAA	.092	2, 108	.025	2.434
12	T-QA, T-A, F-TAA	.003	3,107	.096	4.884

4.4.3.2 Hypothesis 2c: Higher-Mouse-Higher-Objective-Difficulty-Hypothesis

As a preliminary assumption for this hypothesis, objective difficulty and subjective ratings of difficulty should differ. In order to proof this difference, the correlation between objective and subjective difficulty was first calculated. As shown in Table 21, all correlations are negative, indicating that items with objectively higher difficulty (=less often correct) were also rated as subjectively more difficult. However, the correlation coefficients range between -.003 and -.464 with a mean correlation of -.232 which is regarded as small. Therefore, subjective difficulty ratings are significantly different to objective difficulty and, do not seem to be very accurate. Indices of mouse behavior for each BEFKI item during the answering of it, were correlated with the according correctness of the answer. Note that a higher value in the used correctness measure means that the item was less difficult. As for Higher-Mouse-Higher-Subjective-Difficulty (2a), the number of items that have significant positive and negative correlations with indices of mouse behavior was counted and the mean correlation was computed as shown in Table 22. The number of positive correlations (4) is lower than the number of negative correlations (28), indicating that higher objective difficulty (=less correct answers) of items relate to higher indices of mouse behavior (e.g., the higher the total time on the item, the higher its objective difficulty). Moreover, correlations were computed for aggregated indices over all items, ignoring the item level listed in Table 22. Correlations are significant and negative for 4 of 6 indices, but low, ranging between r = -.062 and -.111, as shown in Table 23. Thus, the hypothesis that indices of mouse behavior correlate negatively with objective difficulty is accepted.

Table 21

Correlation of subjective and objective difficulty for BEFKI items ordered by subjective difficulty.

Item	Correctness	SD	Subjective Difficulty	SD	Correlation	p
1	0.97	0.161	0.35	0.778	-0.351	**
2	0.97	0.161	1.13	1.299	-0.111	0.121
8	0.91	0.283	1.16	1.272	-0.256	**
3	0.96	0.185	2.15	1.403	-0.177	0.110
5	0.95	0.224	2.53	1.570	-0.273	**
0	0.78	0.419	2.65	1.430	-0.294	**
7	0.83	0.374	2.70	1.546	-0.057	0.274
11	0.68	0.467	2.95	1.320	-0.321	**
13	0.64	0.482	3.10	1.376	-0.175	*
9	0.78	0.416	3.12	1.548	-0.464	**
4	0.82	0.389	3.19	1.375	-0.248	**
6	0.80	0.403	3.22	1.287	-0.290	**
12	0.70	0.460	3.26	1.534	-0.395	**
10	0.67	0.473	3.55	1.217	-0.281	**
14	0.56	0.498	3.63	1.166	-0.349	**
16	0.47	0.502	3.66	0.988	-0.003	0.487
17	0.24	0.427	3.82	1.054	-0.024	0.400
15	0.36	0.482	4.25	0.892	-0.107	0.130
1	0.97	0.161	0.35	0.778	-0.351	**
Mean	0.73	0.378	2.80	1.281	-0.232	

Note. ** = significant at the 0.01 level, * = significant at the 0.05 level

Table 22

Correlations between indices of mouse behavior on items and objective difficulty.

	T-QA	T-QT	T-A	T-FS	F-A	F-TWA	F-TAA	F-TAQ
Number of items with sign. positive correlation	0	0	1	2	0	1	0	0
Mean r	-	-	.169	.327	-	.196	-	-
Number of items with sign. negative correlation	5	3	4	4	2	2	4	4
Mean r	-.208	-.211	-.261	-.230	-.314	-.263	-.248	-.180

Table 23

Correlations between aggregated indices over all items an objective difficulty.

	T-QA	T-QT	T-A	T-FS	F-A	F-TAA / F-TQA
r	-.069**	.004	-.062**	.007	-.092**	-.111**
p	< .01	.072	< .01	p = .796	< .001	< .001
n	2014	2013	2052	1750	2010	2034

4.4.3.3 Hypothesis 2d: Mouse-Indicates-Objective-Difficulty-Hypothesis

To check whether the mouse behavior indicates objective difficulty of general knowledge items in the BEFKI questionnaire, binary logistic regression models were computed for items with at least two significant indices, entering the significantly correlated indices of mouse behavior as predictors and binary correctness (0 being incorrect, 1 being correct) as the dependent variable. As shown in Table 24, only the models for item 14 and 15 included indices that were significant in the model according to a Wald chi-square test. For item 14, F-TWA, the frequency of transitions between whitespace and answer options was significant, improving the percentage of correct predictions from 55.9% without the variable to 64% after including the variable. For item 15, the time till first selection of the answer was significant, but the beta value was zero, so the index did not impact the regression term.

Considering the low number of significant correlations between items' objective difficulty and indices of mouse behavior, as well as the low performance of mouse indices in the calculated binary logistic regressions, the hypothesis is rejected.

Table 24

Significance of variables for binary logistic regression models with mouse behavior indices as predictors and correctness of the answer as dependent variable.

Item	T-QA	T-Q	F-TQA	T-A	T-FS	F-TWA	F-A	F-A
0	-	$p = .066$	$p = .828$	-	-	-	-	-
1	$p =.930$	-	-	$p = .088$	-	-	-	-
2	$p = .823$	$p = .161$	-	-	$p = .319$	$p = .728$	-	-
3	$p = .489$	-	$p = .448$	-	$p = .989$	$p = .401$	$p = .514$	$p = .374$
4	-	-	$p = .188$	$p = .420$	$p = .322$	$p = .262$	$p = .505$	-
5	-	-	-	$p = .054$	$p = .899$	-	-	$p = .999$
14	$p = .795$	-	-	-	-	$p < .01$, *beta* = -.184		-
15	-	-	-	$p = .985$	$p < .01$, *beta* = 0	-	-	-

4.4.3.4 Hypothesis 3a / 3b: Slower-FOK-Higher-Difficulty / Faster-FOK-for-Extreme-Difficulty

To check these two hypotheses, correlations between the both indices of mouse behavior T-A and T-FS regarding the FOK rating, and the subjective difficulty ratings were computed. The indices were chosen because T-FS represents what is commonly known as the response time for an item (time to first select), and T-A seems to be important as it represents the isolated time on the answer options. As shown in Table 25, all items with a mean FOK greater than .23 either have a positive correlation between subjective difficulty and T-FS, T-A, or both. Mean positive correlations between response times of FOK-ratings and the subjective difficulty were $r = .237$ ($r_{min} = .159$, $r_{max} = .305$) for T-FS, and $r = .342$ ($r_{min} = .245$, $r_{max} = .458$) for T-A. Figure 27 shows T-A and T-FS for every item's FOK ordered by subjective difficulty, as well as the actual subjective difficulty, and FOK. *Slower-FOK-Higher-*

Difficulty-Hypothesis (3a) can be accepted as there is a positive correlation between at least one of the two indices for every item with a FOK of at least .23. *Faster-FOK-for-Extreme-Difficulty-Hypothesis (3b)* can partly be accepted, as one correlation is negative, but for other items with extreme difficulty, no correlation was found (i.e., items 14 - 17).

Table 25

Correlation between subjective difficulty rating and the mouse indices T-FS and T-A for feeling-of-knowledge ratings

Item	T-FS	T-A	Subj. Difficulty	FOK	$r_{T\text{-}FS}$	$r_{T\text{-}A}$	Correctness
15	3603	963	4.25	0.06			0.36
17	3269	853	3.82	0.08			0.24
16	5469	1149	3.66	0.12	-0.164	-0.224	0.47
14	5721	1129	3.63	0.23			0.56
12	5406	1147	3.26	0.46		0.458	0.70
13	3500	652	3.10	0.48	0.300	0.379	0.64
10	7774	933	3.55	0.49	0.212	0.228	0.67
6	7006	1216	3.22	0.53	0.159		0.80
0	3452	663	2.65	0.54		0.250	0.78
4	6727	924	3.19	0.54	0.277	0.368	0.82
11	3729	521	2.95	0.67	0.278	0.343	0.68
9	5139	462	3.12	0.68	0.280	0.380	0.78
5	0	0	2.53	0.75	0.163	0.483	0.95
7	3954	469	2.70	0.75		0.300	0.83
2	5906	647	1.13	0.88	0.229	0.324	0.97
3	3149	251	2.15	0.90	0.230	0.245	0.96
8	5898	437	1.16	0.93	0.180	0.348	0.91
1	4098	290	0.35	0.96	0.305		0.97
MEAN	4929	747	2.80	0.56	0.237[a]	0.342[a]	0.72

Note. [a] = mean of all positive correlations.

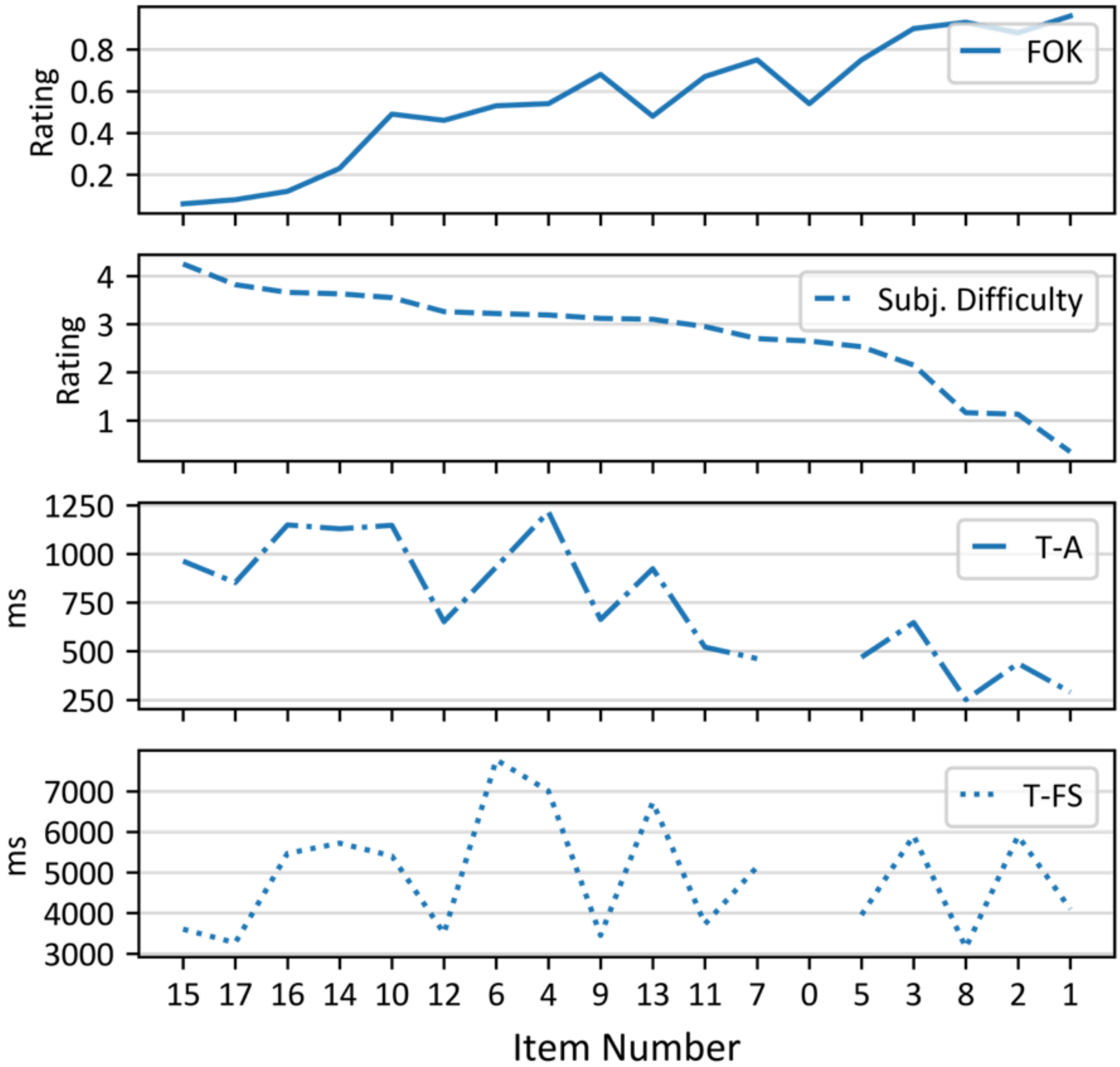


Figure 27. For each BEFKI item: T-A and T-FS during ratings of FOK, FOK-rating, and subjective difficulty. Items are ordered by subjective difficulty.

4.4.4 Discussion

Detecting and Measuring Confusion Through Mouse Behavior

In the first question of this study, the lack of an applicable and unobtrusive method to detect and measure confusion was addressed. Questionnaires were chosen as ecologically valid materials that are often used in learning environments. For the first *Mouse-detects-confusion-Hypothesis (1a)*, it was successfully shown that manipulated items, which induce confusion through contradictory or grammatically wrong questions, lead to higher indices of mouse behavior than non-manipulated items. In a first step, time-related indices of mouse movements were identified to reliably be higher for items with induced confusion. It is only for the item with the lowest qualitative induction of confusion via grammar errors that non-manipulated items exceed the values. This item was also not rated as much confusing as the other items in the pre-test. The findings are substantial, as five of six manipulated items could be identified by just looking at the highest values of the mentioned indices. In contrast, frequency-related indices such as the number of transitions between the questions and an answer option were not found to tell about the manipulation. On the one hand, this seems to be surprising, as Horwitz and colleagues (Horwitz, Kreuter et al., 2017) identified the number of transitions as significant in predictive models. On the other hand, the studies are hardly comparable in this aspect as the authors looked at correspondence to the subjective difficulty rather than confusion. Moreover, they did not use a differentiated index for transitions, but more coarse measures including F-TWA, F-TAA, and F-TAQ. Thus, they might also have had higher statistical power.

In order to visualize the extend of the differences that were found, the effect sizes for differences between T-QA of all manipulated an non-manipulated items were computed. The index was chosen as it corresponded to 5 of 6 items and does only include mouse movements (no clicks). The adapted Cohen's *d*, that was used to account for different standard deviations as an effect size, range between medium (d = .568) and very high (d = 3.671). Which index of mouse behavior to choose is not trivial, and this study contributed to identify which are suitable to be used for this application. In order to reach more explanatory power, a conglomerate was also used to show that manipulated items can be detected by the respective mouse behavior. The conglomerate combines all time-related indices, and as indices were not collinear, no index was left out. Using this conglomerate, it was shown that all manipulated items were significantly higher than the overall-mean of all items on a page with 10 items. However, the K-value of item 39 on page 4 was non-significant, but still slightly higher than for the item of the manipulated item 38. A possible explanation for this might, at the same time, be an interesting phenomenon, that can be observed for manipulated items of other pages as well (see Figure 24). K-values for items that are positioned after a manipulated item seem to also have increased indices of mouse behavior, as observable for item 4, item 18, item 24, and item 48. It is argued that the manipulated, confusing item interrupted the interaction so strongly, that it takes some time for participants to fade out. Increased attention of participants to discover other contradictions or grammar errors in the following items could be the reason for these higher values. Hence, it is possible that the confusion induced with item 38 led to the higher values of item 39. Although a reverberation effect must still be confirmed, a possible intervention could be derived from it. If an instructional designer wants more attention on a crucial item, a confusing item could be placed just before the item of interest – where an item does not necessarily need to be a survey item but might also be another task. This is similar to the idea of perceptual and conceptual disfluency, which was found to increase metacognitive activity such as judgements of learning (Schwarz, 2010).

In the second part of the first hypothesis (*1b*: *Mouse-measures-confusion-levels-Hypothesis*), different manipulations that were meant to induce different levels of confusion were compared. It could be shown that items with contradictions triggered significantly higher indices of mouse behavior compared to items with grammar errors, with a medium effect size of d = .568. Moreover, grammar items still accounted for significantly higher indices than non-manipulated items, with a large effect size of d = .904. Although this result seems to be intuitive, it is very meaningful for the unobtrusive measurement of confusion. It is a theoretical and methodological challenge to not only detect confusion, but also measure different levels of confusion (e.g., Arguel et al., 2017). Notably, the different levels could be identified from the conglomerate that represents the number of participants who showed higher indices of interaction, but also for the isolated indices that represent continuous, individual variables.

Relations Between Mouse Behavior and Item Difficulty

The second research question of this study asked whether mouse behavior corresponds with the self-rated, subjective and objective difficulty of general knowledge items. *Higher-Mouse-Higher-Subjective-Difficulty* (*2a*) stated that indices of mouse behavior correlate positively with subjective difficulty ratings. On an item level, more indices were related positively with subjective difficulty than negatively. When ignoring the item level, all indices correlated low but significantly positive with subjective difficulty. This is in line with previous research. Horwitz and colleagues (2017) computed the average number of mouse movements separately for every step of the Likert-scaled subjective difficulty rating. They found "a significant increase in the number of movements, as participants reported more difficulty answering". (Horwitz, Kreuter et al., 2017, p. 10). These results could be replicated in this study with minor modifications. Instead of the number of mouse movements, detailed indices of mouse behavior were used. All used indices are significantly different between the levels of reported subjective difficulty. A relationship between increasing difficulty and the level of mouse behavior indices can be clearly seen, but the relationships are not always reliable: for three of the five indices (T-QA, T-QT and T-FS), the indices for the highest reported difficulty ("very difficult") are smaller than for the previous one. For index F-A (number of answer selections), "rather difficult" is below "medium". The "number of mouse movements" used by Horwitz and colleagues (2017) is closest to the transition indices used in this study. The relationship between the combined transitions (F-TAA and F-TQA) and the reported difficulty is very similar to the results of the mentioned study.

Although the results of *Higher-Mouse-Higher-Subjective-Difficulty-Hypothesis (2a)* indicate, that higher indices of mouse behavior are related to higher subjective difficulty, the idea of using mouse behavior as a substitute for subjective difficulty ratings (*Mouse-indicates-subjective-difficulty-Hypothesis, 2b*) could not be reliably applied according to the results of the multiple regression models that were computed for different items that have significant correlations with indices of mouse behavior. In this study, the explained variance was not high enough to replace subjective difficulty ratings with index of the mouse behavior. Although it is very promising, that the number of mouse movement transitions between a question /answer and another answer is increasing with the rated difficulty, this needs to be shown on a single item level in order to be used in educational settings or the quality assessment of questionnaires. It is also notable that the frequency indices do provide information regarding the subjective difficulty, but, as mentioned in the discussion of the previous hypothesis, not regarding the confusion level.

As another part of the second research question, the relation between mouse behavior and objective difficulty, operationalized as an answer being correct or not, was investigated. As a preliminary assumption, it was first shown that the difficulty ratings correlate with the actual correctness, but that correlations were not very high - a result that is in line with research on metacomprehension on the judgement accuracy of learners (e.g., Dunlosky & Lipko, 2007). Correlations between mouse indices and objective difficulty do exist both on an item-level and on an overall level (*Higher-Mouse-Higher-Objective-Difficulty-Hypothesis, 2c*). Although statistical significance was reached due to the high power of the correlation on an overall item level, these correlations are very low. Looking at the binary regression model, indices of mouse behavior do not seem to be a proxy measure for objective difficulty in this study, which was contrary to the expected *Mouse-indicates-objective-difficulty-Hypothesis,* *2d*. Unfortunately, there is no other study yet that investigates this relation and that these results can be compared with.

Relations Between Mouse Behavior on FOK and Subjective Difficulty

In the last research question, it was investigated whether the response time for metacognitive FOK-judgements, operationalized by indices of mouse behavior, are related to the subjective difficulty of general knowledge items. In this research question, the interaction with the FOK ratings was investigated, not the interaction with the actual BEFKI questions. As hypothesized, participants showed higher response times at their FOK ratings when rating a question that they perceive as more difficult.

The operationalization of response time with indices of mouse behavior is crucial in this study. Traditional response times are only a coarse estimate based on presenting one item at a time on a page and taking the time that a participant spends on this page. More sophisticated approaches look at the time between two answers when multi items are presented on one page, called T-FS (time to first selection of a question) in this study. In addition to this, this study looked at the index T-A, which is the time that participants spend on the answer options with their mouse pointers. Correlations with subjective difficulty are higher for T-A than for T-FS, indicating that T-A is a better indicator for subjective difficulty (*Slower-FOK-Higher-Difficulty-Hypothesis, 3a*). This is also apparent from Figure 27, as it indicates how T-A declines with items getting less difficult.

The second hypothesis of this question (*Faster-FOK-for-Extreme-Difficulty-Hypothesis, 3b*) predicted an opposite picture for very difficult items. Among the four most difficult items, however, a negative correlation between the mouse behavior during the FOK ratings and the subjective difficulty of the item was found only for one item. The other items did not show a negative correlation, but no positive correlation either. The absence of a correlation therefore at least does not contradict the theoretical assumption of Glucksberg and McCloskey (1981) that don't-know ratings are made faster if no relevant knowledge is available.

The results of research question 3 are very promising, because the interaction with a very simple binary question, whether a person knows the answer to a question, corresponds to how difficult the person perceives this question. Although this might be an intuitive relationship, it has not yet been empirically investigated and is, by no means, self-evident.

4.4.4.1 General Limitations

There are some general limitations to the results of this study. An issue in many studies that try to replace subjective self-reports with objective, unobtrusive measures is its circular argumentation: the aligned measure of confusion (in this study, mouse behavior) can only be valid if the self-report on confusion is valid. Moreover, the rating during the study could be reactive to the actual behavior with the given item. These issues are inherent to the research paradigm. Hence, this study tried not to validate the confusion with self-reports during the task. Instead, items that were designed to induce confusion were pre-tested to validate the induction and the level of induced confusion. The methodology to rate the items could have suffered from low ecological validity. Rating questions regarding the level of confusion they induce is not a common task. Thus, the task might rather operationalize whether the manipulations in the items could be identified. On the other hand, the pre-test successfully identified different levels of confusion for grammar errors and contradictions using the full range of the Likert-scale. Hence, there was no ceiling-effect for manipulated items in general.

The same issue occurred in the research questions that deal with the general knowledge questions. The findings can only be valid if the rating of the subjective difficulty and the FOK was valid. Moreover, the previous rating of subjective difficulty and FOK can already be reactive to the actual answer and the mouse interaction with an question and it’s answer options. This issue was tried to be addressed by presenting different questionnaires between the item rating and the actual answering of the items.

Another, methodological drawback lies in the computation of the conglomerate K. This index represents the number of cases right of the median of all indices and thus, is accompanied by a loss of information: it only considers whether the values of the individual participants are higher than the median, but not to what extent they deviate.

4.4.4.2 Application

An application of the findings would be the use of data on mouse behavior regarding the answering of rapid assessment tasks (e.g., Renkl et al., 2015). Considering not only the correctness of the answer, but also the mentioned indices of mouse behavior could contribute to more accurate learner models in adaptive learning environments.

The results could also have a completely different framing, not looking at how to recognize the state of confusion of a learner, but recognize content that induce confusion. This content can potentially be learning materials, but also the very same material that was used in this study: questionnaires. The use of indices of mouse behavior could act as a new measure for the validation of items for pre-testing questionnaires and scales. As shown, such measures can identify items that induce confusion.

4.4.4.3 Conclusion

Multi-item scales provide an inherently suitable environment to find the effects that were looked for, as manipulated (confusing) and non-manipulated (non-confusing) items differ in their content, but their structure remains the same. Although this was an important first step to proof that it is possible to recognize confusion by mouse behavior, a transfer of these results into other environments that are more usual in learning contexts and thus, not as structured as multi-scale items is needed. Regarding the research on recognizing difficulty in multi-item scales, this study could contribute by confirming and extending the findings that mouse behavior can indeed contribute to measures of subjective and objective difficulty.

Moreover, the results of this study also contribute to the general discourse on confusion as an academic emotion, claiming that the pausing behavior is a response to the confusion that was induced. Less granular measures that have been used (e.g., time on page or item response time) in learner models cannot tell what the source of an changed value is. Using measures with higher granularity on element-basis instead of page-basis allows for a more detailed analysis of what causes higher / lower values.

4.5 Intervention Study: Can Metacognitive Prompts Boost the Effects of a Learning Dashboard?

In chapter 2.5.3 of this work, learning dashboards were introduced as a recent instructional intervention that supports SRL processes. Moreover, the current drawbacks were listed. The goal of this study is to incorporate recommendations of the recent reviews that have been introduced and that address theoretical and empirical issues of dashboards. Specifically, the following issues will be addressed in this study:

1. General mechanisms of how dashboards impact learning are often not stated or not based on theories of educational psychology, and
2. data channels and its visualization presented in dashboards do not account for theories of educational and cognitive psychology (Gašević et al., 2015).
3. Simply raising awareness does not seem to be enough to facilitate learning, and studies that report positive effects on learning through learning dashboards are very rare (Jivet et al., 2017).

4. There is a general lack of systematic experimental research on the effect of dashboards on learning performance and on the use of dashboards by learners (Bodily & Verbert, 2017; Jivet et al., 2017; Schwendimann et al., 2017)

Regarding 1): We examine the use of a learning dashboard and the effects on learning performance from the theoretical perspective of SRL. From the range of existing frameworks, the COPES model was chosen that describes five facets (condition, operation, product, evaluation and standard; Winne & Hadwin, 1998) of four stages (task definition, goal setting and planning, enacting study tactics and strategies, and metacognitively adapting studying) that build a “recursive, weakly sequenced system” (Winne & Hadwin, 1998, p. 281) for learning processes. A simple reason for using the COPES model in the few dashboard studies that use a theoretical framework, it is the most prominent. Using COPES makes this study more comparable to existing and future studies. Another, even more important reason is, that COPES allows us to explicitly spot the underlying mechanism of the intervention as follows.

According to COPES, knowledge or skill acquisition happens through “enacting study tactics or strategies” during the third stage. However, it is argued that rather than directly triggering tactics / strategies on a cognitive level, dashboards take effect on a metacognitive level which can subsequently lead to changes in applying learning strategies. Learning dashboards (should) provide learners with objective information about their learning process that act as sources for external evaluations on task conditions (e.g., time or resources) and cognitive conditions (e.g., domain knowledge), that allow for better cognitive evaluations (e.g., “Am I on target with this task?”). As described in the COPES model (see chapter 2.2), if learners act on (valid) evaluations, they can adapt their standards, change their cognitive conditions, and triggers new operations (“If the student acts on evaluations, this is control by which elements in the collage of cognitive conditions may be altered; standards may be adjusted, added, or abandoned; and, operations of new kinds may be carried out.”, Winne & Hadwin, 1998, p.

281). This process is depicted in the blue parts of Figure 28. As stated, learners need to act on these updated evaluations in order to positively change learning and learning outcome. This brings us to the second claim.

Regarding 2), the dashboard in this study uses fine-grained interaction data that is presented to learners using different visualizations for different purposes. Simple column charts (vertical) and bar charts (horizontal) have been used whenever possible, as previous research (Simkin & Hastie, 1987) has shown that they lead to the highest accuracy in data interpretation compared to other visualizations such as pie or line charts (when not used to compare proportions of the whole, where pie charts performed best). This was the case for the visualization of current task status and navigation. However, heat maps of mouse positions were used to provide learners with detailed information on what content the interacted with or not. The rationale behind presenting heat maps of mouse movements is that it should help learners with updating their cognitive evaluations on a detailed, element level. It is argued that heat maps indicate such information as there are medium correlations between mouse movements and gaze behavior (e.g., Guo & Agichtein, 2010; Huang & White, 2012). Moreover, there is also a line of research called eye movement modeling examples (EMME, see chapter 2.5.4) that uses previously recorded eye movements to foster learners cognitive processing of text and images (e.g., Mason et al., 2015). Although EMMEs use eye tracking information of others processing the information in an efficient way, the fact that it fosters processing is an indicator that learners should be able to gather information from such visualizations. The argumentation is that not only expert models could help improving, but also the comparison to learners' own interactions.

However, as heat maps are rather used when it comes to the visualization of focused areas in eye tracking for analytics instead of interventions (see Špakov & Miniotas, 2007 for an introduction), there is still a lack of research on the information processing from heat map visualizations. Hence, this study may also shed light on how able learners are to handle heat maps as a visualization and source for inferences on their learning. The theoretical reasons why the specific data sources were used with regards to the COPES model are explained in the methods section where the intervention is described in more detail.

Regarding 3) Simply raising awareness through learning dashboards does not seem to be enough to foster learning. Prerequisites for “raising awareness” in the sense of updating cognitive conditions and evaluations of learners as mentioned in the COPES model are that 1) correct information about the learning process is available to learners, and 2) they correctly process this information. Although some studies focused on these prerequisites, there is no clear evidence regarding what data sources are best suitable to achieve that (Jivet et al., 2017). Even if this was fulfilled, learners might not act on resulting evaluations. It is argued that metacognitive prompting could help learners to incorporate the dashboard information in their further learning process. The so-called “production deficit” describes the reason why such prompts could work: although learners are often skilled and have previously acquired learning strategies, they do not use them spontaneously (see chapter 2.5.1). This concept can be transferred to the intervention of this study, as the strategies remain the same, but are used with different parameters (task conditions, cognitive evaluations) that have been updated with the help of information from the learning dashboard. Orange parts of Figure 28 indicate how metacognitive prompting affects learning in the COPES model.

Regarding 4) A general lack of systematic experimental research is counteracted by applying a rigorous 2x2 factorial design including a control group. Moreover, the interaction with dashboards are examined on different conceptual levels:

- Firstly, based on the software framework ScreenAlytics (see chapter 4.1), fine-grained interaction is recorded through peripheral data to shed light on how learners are using the intervention.

- Secondly, CL is assessed that is induced by the dashboard.
- Thirdly, learners are asked about the subjective usefulness of the different parts of the learning dashboard regarding the support of their learning process.

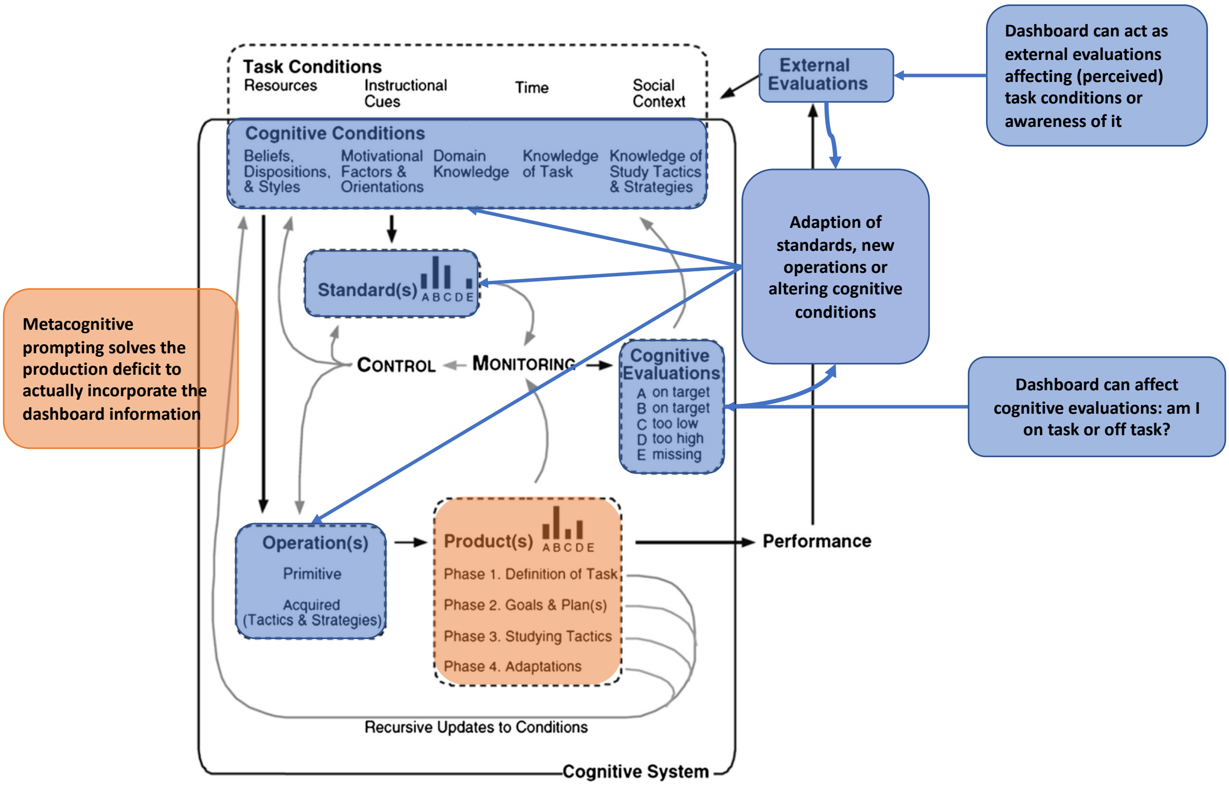


Figure 28. Adapted from Winne and Hadwin (1998). Blue parts (right) indicate how dashboard affects cognitive and external evaluations. Orange parts (left) indicate how metacognitive prompts affects the products during the learning process.

4.5.1 Research Question and Hypotheses

The aim of this study is to experimentally examine the effects of the two interventions “learning dashboard” and “metacognitive prompt”, and a combination of both on the learning outcome in contrast to a control group. The following hypotheses are therefore stated:

1. *Prompts-And-Dashboard-Hypothesis:* Learners who receive learning dashboards paired with prompts will have higher learning outcomes compared to learners who receive prompts only, learning dashboards only or no intervention.

2. *Prompts-Or-Dashboard-Hypothesis:* Learners who receive learning dashboards or prompts will have higher learning outcomes compared to learners who receive no intervention.

Non-linear navigation behavior is, according to Astleitner (1997), an indicator of systematic learning behavior, since students consciously and purposefully decide which nodes should be selected. Moreover, it has recently also been empirically shown that prompting leads to more non-linear navigation (Pieger & Bannert, 2018) and is connected to better learning outcomes (Bannert et al., 2015). Thus, it is argued that non-linear navigation behavior in this study, can be a meaningful representation of better evaluations induced through the interventions, especially through the dashboard. Therefore, the following hypothesis is formulated.

3. *More-Non-Linear-Navigation-Hypothesis:* Learners that receive an intervention show more non-linear navigation behavior than learners who do not receive an intervention.

Furthermore, as there is a lack of knowledge about how learners use these interventions (e.g., Bannert & Mengelkamp, 2013 for prompting), the interaction between the learners and these interventions is examined including detailed data of the usage of different parts of prompts and dashboards, the induced CL through the learning dashboard, and the perceived usefulness of the interventions by the learners. This leads to the following explorative questions:

1) How do learners interact with prompts and the learning dashboard?
    a. How long and frequently do learners interact with the interventions?
    b. What CL is induced by the learning dashboard?
2) How do learners perceive the learning dashboard?
    a. Do learners perceive the dashboard as useful for their learning?
    b. Which parts of do learners perceive as useful, and which not?

4.5.2 Method

4.5.2.1 Sample and Design

This study was conducted as an online field experiment with a 2 (prompt vs. no-prompt) x 2 (dashboard vs. no-dashboard) factorial pre-post between-subject design. For participant recruiting, an advertisement was posted in seven Facebook groups of different German universities that were addressed to first semester students. The advertisement involved the following information: participation takes approximately 60 to 75 minutes, the task is to "test a learning environment about programming, work on short quizzes and fill questionnaires", requirements to participate are a calm environment, a laptop or desktop computer (no tablets or mobile phones), no to very low prior knowledge about programming, 10 Euro will be payed as a compensation for a complete participation. Participants were randomly assigned to one of the four groups. Out of 209 participants, 138 completed the study and fulfilled the following requirements to get included into the further analysis: 1) self-reported no or low prior knowledge, 2) a checkbox with the label "I want to seriously take part and finish in this study" was checked, 3) all questionnaires and quizzes were completed, 4) a desktop computer or laptop was used, 5) a meaningful interaction with the environment was visible on all accessed pages in the screen recordings in order to prevent clicking-through. Page interactions were rated as meaningful if there was mouse movement and scrolling for at least 30 seconds.

Participants were between 18 and 34 years old (M = 21.34, SD = 2.80), 44 were male and 94 were female. Random assignment of the participants to groups resulted in the following distribution: 37 in the control group, 37 in the prompt-only group, 31 in the dashboard group and 33 in the prompt+dashboard group.

4.5.2.2 Learning Materials

Students had to learn the basic concepts and terms (e.g., loops, functions, variables), and the syntax (e.g., where to place brackets and semicolons) of JavaScript, a common programming language used mainly in web applications. The material consisted of 16 content pages including about 3000 words, three tables, two illustrations, ten code examples and five interactive coding exercises. Students could run code examples by clicking on a "Try" to see the results of it, as shown in Figure 29.

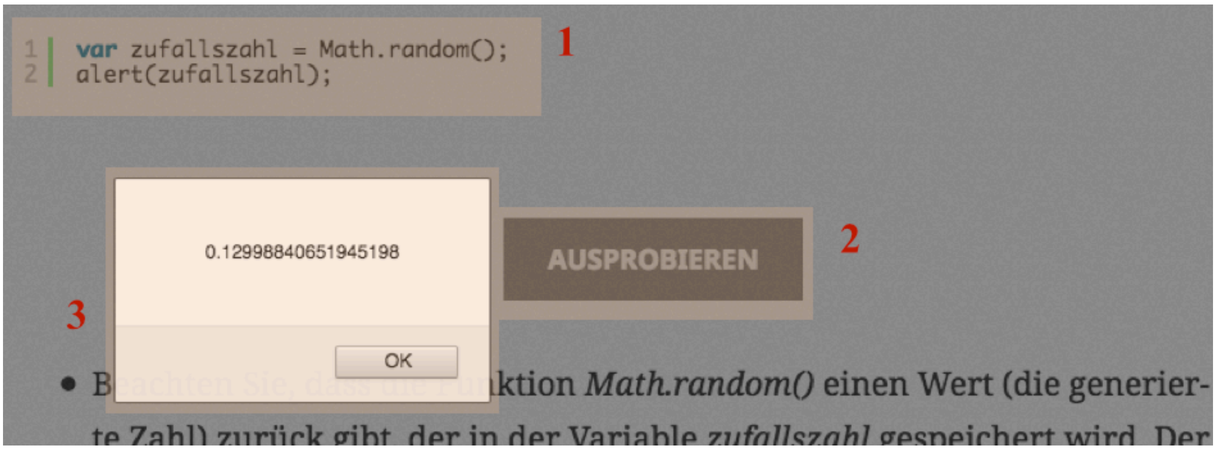


Figure 29. Students first studied the code given in area 1, then clicked on the button 2 and received the result of the code in window 3.

In interactive coding exercises, learners had to write their own JavaScript code to solve a given task (e.g., "Assign the following values to the given variables."). A pedagogical agent, introduced as "Anna", gave feedback and automatically recognized mistakes in the code. Students saw the result of their codes after clicking on a "Try" button. Figure 30 shows an example of a coding exercise. Although the learning material was structured linear, navigation back and forth was possible through a menu on the left. Learners were provided with a function to take notes and the remaining time was shown on the left (see Figure 31).

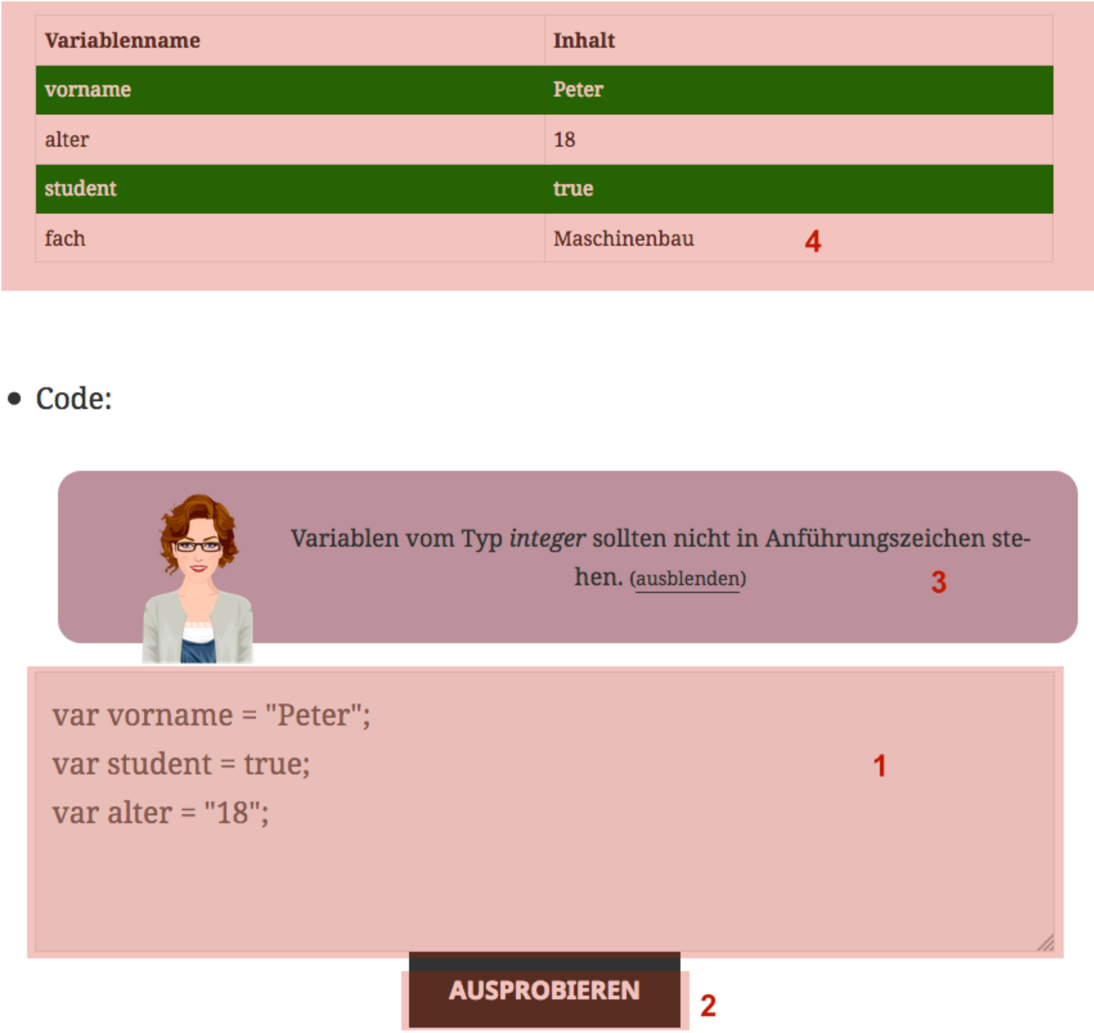


Figure 30. In coding exercises, students wrote their own code in a text box (1) to solve a given task. After clicking on the "Try" button (2), the pedagogical agent (3) gave feedback on mistakes. Green rows in the table above (4) indicated correct solutions.

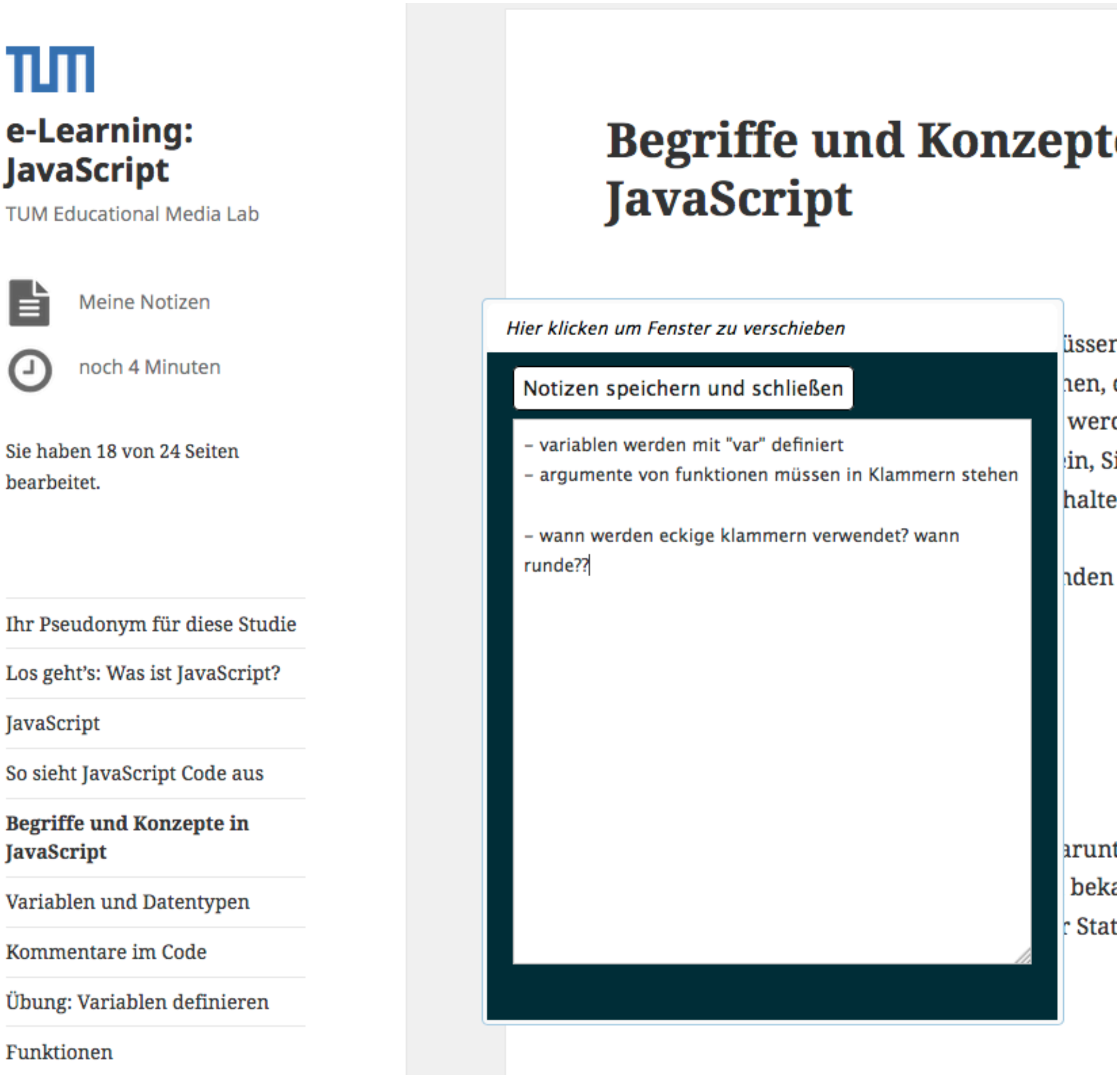


Figure 31. In the learning environment, learners could navigate through a menu on the left and could use a window to take notes. The remaining learning time was presented on the top left.

4.5.2.1 Interventions: Dashboard and Prompt

The intervention of this study consists of two parts that were presented to the different experimental groups combined and separately: metacognitive prompts and the learning dashboard. All groups received the intervention two times, after finishing approximately one-third and two-third of the learning materials. Both groups, prompt and prompt+dashboard, got the same metacognitive prompts, but learners in the prompt+dashboard group got additional instructions to review the presented visuals in the dashboard (see Table 26 for the texts).

Prompts were displayed in a popup window on a blank page for the prompt group, or on the dashboard page for the prompt+dashboard group. Learners needed to click “OK” to close the window (see Figure 32) and the prompt group was redirected to the learning environment after closing the window.

Table 26

Texts of Metacognitive Prompts for the Prompt and Prompt+Dashboard Groups

Group	Metacognitive Prompt Contents
Prompt	Before you continue with learning, please take some time to think about the following questions: 1) Which content and pages do I already understand well and which not? 2) Which pages should I study again? 3) What can I do to clarify the things that I don’t understand yet? 4) What should I change in the way I’m currently learning?
Prompt+Dashboard	Before you continue with learning, please take some time to study the visuals regarding your learning process. While you do so, ask yourself the following questions: 1) Which content and pages do I already understand well and which not? 2) Which pages should I study again? 3) What can I do to clarify the things that I don’t understand yet? 4) What should I change in the way I’m currently learning? You can review these questions by clicking on the associated tab. Click “back to the learning environment” to continue with learning.

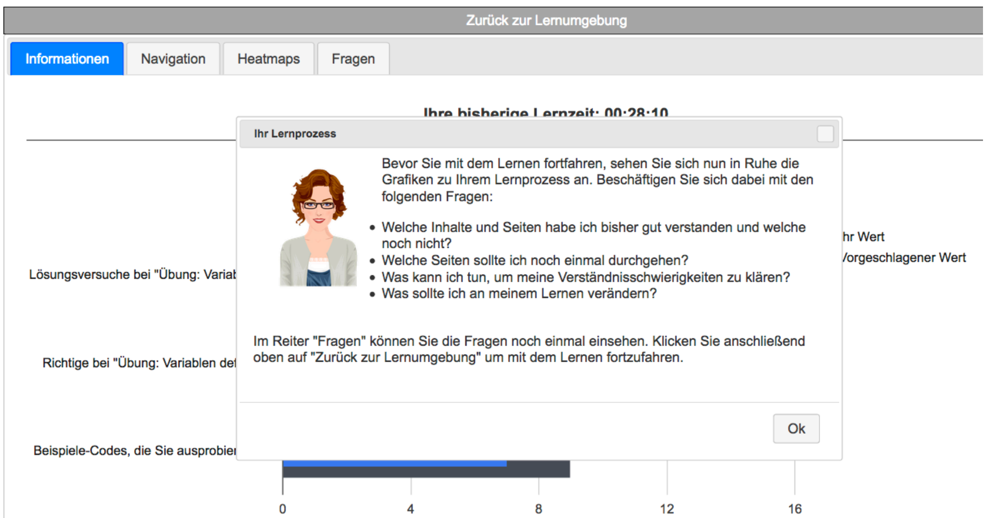


Figure 32. Metacognitive Prompts Were Shown in a Popup-Window on the Dashboard Page.

As shown in Figure 33, the dashboard contained four different tabs that learners could switch between: “Information” contained the learning time and the current status of exercises (number of solution attempts, number or percentage of correctly solved steps, number of viewed example codes) in separate bar charts. Learner’s own values were shown in blue bars, while a “suggested value” for each metric was shown in an underlying grey bar as references of what should be reached. For exercises and tasks, these were set to the maximum possible score, for the number of solution attempts, the number of sub-problems was counted of the according exercise and for number of viewed example codes, the number of available example codes was taken, so that each should be viewed at least once. Within the COPES model, this information should provide resources to update the standards through of learners. In the “Navigation” tab, a visualization of the path that the learner took through the

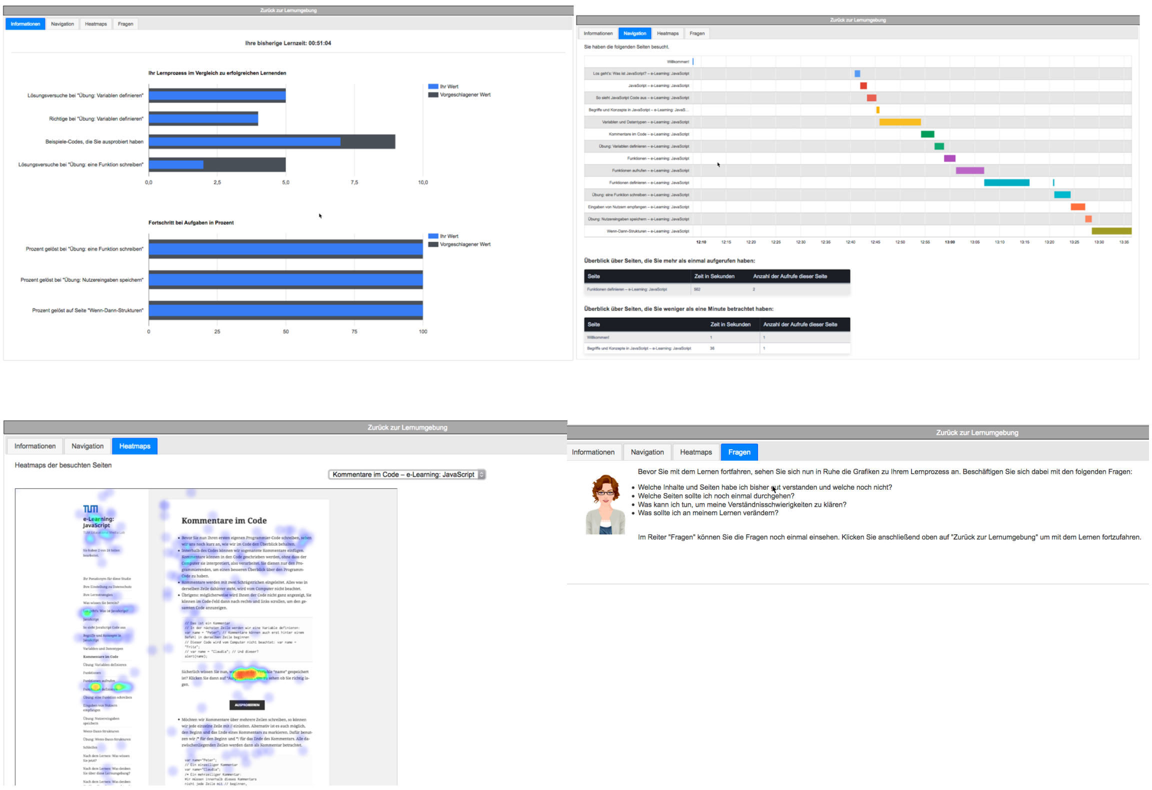


Figure 33. Available Tabs in the Dashboard: Information (top left), Navigation (top right), Heat maps (bottom left) and Questions (bottom right).

learning environment and the view times of each page was presented. Moreover, a list indicated pages that were viewed less than one minute and pages that were viewed more than once. Referring to COPES, this aims at updating learners' task conditions, specifically the remaining and spent time. The tab "Heat maps" contained a visualization of the mouse movements on a page (as described in chapter 4.1.3.3). The viewed page that the heat map should be shown for was selectable via a dropdown menu. In contrast to the time, this was implemented to provide more detailed information on what content resources have already been used and which not. The argumentation for the use of mouse movements to infer which elements have been used or looked at is based on correlations between eye gazing and mouse movements, and between eye movements and attention (de Koning, Tabbers, Rikers & Paas, 2010; Huang & White, 2012). Moreover, in a mouse-controlled learning environment, mouse movement does inherently tell what content has been used. As an example, if a button to run a simulation in a learning environment has not been clicked with the mouse, it has definitely not been used by the learner, which would be visible in such a heat map.

The "Questions" tab was only available for the prompt+dashboard group and listed the questions initially shown in the metacognitive prompt to enable learners to review the questions. Above the tabs, a button saying "back to the learning environment" was placed.

4.5.2.2 Procedure

The procedure of this study is depicted in Figure 34. On the initial page of the experiment, participants were instructed about requirements to participate (device, time, calm environment, no to low prior knowledge), payment procedure of the incentives, a short description of the learning materials, data that is being collected in the study and that is connected to their pseudonym. Bank account information was deleted immediately after checking the requirements and transferring the incentive money in order to ensure data privacy of the participants.

After the initial page, participants were randomly assigned to a group and presented with a video training on how to use the learning environment. Participants in the prompt, dashboard and prompt+dashboard groups were presented with additional instructions on how to use these interventions. The videos included a pedagogical agent that presented the following content: 1) Introduction of the pedagogical agent "Anna", 2) description of the learning contents, 3) instructions on how to use the menu and notes function, 4) instruction that no other materials than this website should be used, 5) instruction on how to view the remaining time and the percentage of completed content. Additionally, according to the assigned experimental group, participants received an introduction to the effects and use of metacognitive prompting and/or learning dashboards as this was found to be necessary in order to ensure the effectiveness of such interventions (e.g., Bannert et al., 2015). The control group received an introduction to ergonomics (workplace design, e.g., how to adjust the computer monitor and chairs) as used in Bannert, Sonnenberg, Mengelkamp & Pieger (2015) in order to maintain an equal workload. This alternative training was not related to the learning contents and activities.

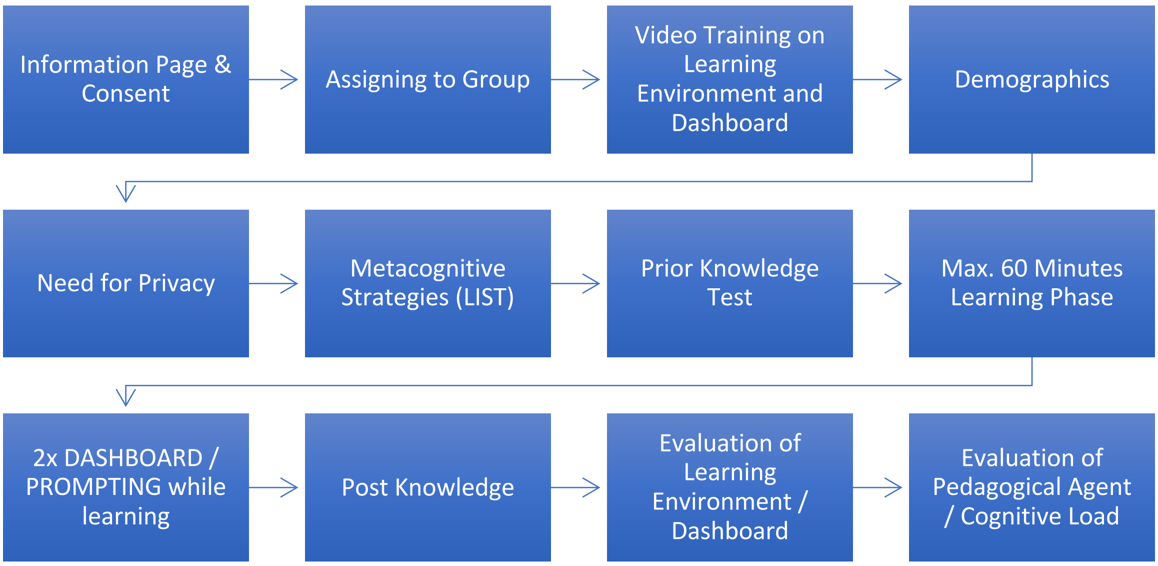


Figure 34. Procedure of Study 4

After the instructional videos, demographic variables (pseudonym, age, sex, occupation, prior experience with programming, willingness to seriously complete the study), a self-constructed questionnaire on the need for privacy, an adapted version of the LIST questionnaire to assess existing self-regulation strategies, and a self-constructed test on declarative and procedural knowledge to assess prior knowledge were presented. Participants than had a maximum of 60 minutes to learn the presented contents but could also finish earlier. This was done in order to keep a high ecological validity. It was not possible for learners to skip an unviewed page. However, once a page has been visited, learners could move back and forth. After 10 and after 18 pages, the prompt and/or dashboard intervention was presented according to the assigned group. The positions were selected as they represented approximately one-third and two-third of the overall learning material. There was no intervention for the control group, but the time that learners spent on the intervention was not taken from the maximum learning time of 60 minutes. Learners were warned on the last content page that after moving beyond, they will not be able to get back to the learning contents to prevent unintentionally quitting the learning process.

After finishing the study of the learning contents, the knowledge test was presented again as a post measure for learning outcomes. A self-constructed evaluation questionnaire of the learning environment and the pedagogical agent followed for the control group and the prompt-only group. The dashboard and prompt+dashboard groups were presented with an evaluation of the dashboard and a self-report measure of CL induced by the dashboard instead. After completing all questionnaires, the participants were asked to provide their bank account information in order to receive their incentive of 10 Euro.

4.5.2.3 Measures and Instruments

For each of the used scales, the number of valid data rows, number of items, minimum value, maximum value, mean value, standard deviation, and Cronbach's Alpha is listed in Table 29.

Peripheral Data

We used ScreenAlytics (see chapter 4.1) to collect data about the interaction between the learners and the learning environment (such as navigation data, web content, mouse behavior and performance in different tasks, e.g., percentage of solved problems in a coding task) for different reasons. First, the data was the foundation of the intervention, displaying the collected data in order to inform learners about their processes. Second, the acquired data delivers important information on how learners used the intervention (e.g., how long does a learner actually use a prompt) which relates to the explorative research question 2a) and b). Third, as the data collection was in the field, recorded data was used to identify learners who showed usage patterns that did not comply with the requirements for a valid participation (such as clicking through the experiment without reading, as described in chapter 4.1.5.1).

Need for Data Privacy Protection

Data privacy protection (DPP) became an intensively discussed topic in public policy, but also in the field of learning analytics (Drachsler & Greller, 2016). Moreover, other studies report DPP concerns when recording the screens of users (e.g., Tang et al., 2006). Hence, it seems to be of high relevance for the learners in this study and the individual attitude towards data collection could have an effect on the use of the dashboard intervention. If there is an aversion to data collection, students might have a negative attitude towards using the learning dashboard – and the other way around. As there was no existing questionnaire that measures the need for DPP at the time of planning this study, a 10-item scale was developed to assess it. Among others, a possible definition that fits the dashboard application in this study is privacy as “the right of a person to determine which personal information about himself/herself may be communicated to others” (e.g., Walters, 2001, p. 151). Based on this definition, a 5-step (“strongly disagree” to “strongly agree”) Likert scale was developed that asks for the attitude towards DPP and actions that users take in order to achieve protection. Based on the collected data, an explorative factorial analysis was conducted with Varimax rotation method to find the components that items are loading on. Eigenvalue threshold of the components was set to 1. Item 1 was recoded so that higher values meant a higher need for DPP in all items.

Table 27 shows the correlations between the items and the extracted factors sorted by from highest to lowest.

Table 27

Eigenvalues and item-factor correlations for the factors extracted by an explorative factorial analysis with Varimax rotation for the self-report scale of need for data privacy protection.

Item	Factor 1 (EV = 2.46)	Factor 2 (EV = 1.34)	Factor 3 (EV = 1.03)
2 - Data protection in general is an important topic to me.	0.724	0.129	-0.067
4 - When using an app, I always check the privacy settings	0.696	0.102	0.163
7 - I try to protect myself from data abuse (e.g., by using encrypted messengers, firewalls, deleting my cookies, using a proxy server)	0.683	-0.111	0.438
1 - As long as a service is comfortable to use, I don't really care about privacy.	0.074	-0.304	-0.585
6 - It worries me that I leave traces in the internet.	0.154	0.665	-0.129
5 - I think that I do not have enough technical knowledge to take care of my data privacy protection.	-0.136	0.600	0.213
8 - It worries me that my personal digital data could be read by others.	0.037	0.593	0.275
10 - I try to avoid using my real name in the internet.	0.382	0.588	0.069
9 - I'm trying to avoid using devices that collect data about me (e.g., fitness trackers)	0.165	-0.024	0.661
3 - I don't install an app if it asks for too much personal data.	0.137	0.110	0.558

Note. For each item, the highest Eigenvalue is marked. Numbers in brackets after the item copy indicate the position of the item in the presented questionnaire.

Item 1 was removed as it did not load on an extracted factor. Items 9 and 3 were removed as they loaded on a factor that had a very low Eigenvalue of just above 1 (1.03) and did not load high on another factor. As factorial analysis is a method to reduce data, only items were kept that explain more variance then the original variable, meaning their Eigenvalue is above 1 (Kaiser-Guttmann criterion, see Guttman, 1954).

The extracted factors seem to represent different knowledge and competencies of handling DPP. While the actions described in the items of factor 1 describe a general interest in DPP and require a technical understanding (e.g., Item 4, using firewalls or proxy-servers for protection or item 7, checking the privacy settings in an app), the items of factor 2 describe the anxiety of persons and low-level actions (e.g., Item 10, not using my real name in the internet) that could be caused by a low technical understanding (i.e., item 5, not having enough technical knowledge to protect myself). Cronbach's Alpha was calculated for internal consistency for factor 1 (α = .60) and factor 2 (α =.53) and for all seven items (α = .60). A mean of all remaining items was computed as a score for the need for DPP.

Metacognitive Strategies Adapted from LIST

Three subscales of the "Inventar zur Erfassung von Lernstrategien im Studium" (LIST, Schiefele & Wild, 1994) were used, each having 4 items, to assess metacognitive strategies: metacognitive planning (e.g., "Prior to learning, I think about how to learn most effectively"), monitoring (e.g., "I ask myself questions about the topic to ensure I understood everything correctly."), and regulation (e.g., "I adapt my learning techniques if I have to read a difficult text"). As described in the LIST validation study (Schiefele & Wild, 1994), the mean value was used of the mentioned items as an indicator for metacognitive strategies of the participants in this study. Internal consistency, computed as Cronbach's Alpha, was lower in this study (Cronbach's α = .46) than in the validation study of the scale (Cronbach's α = .64).

Declarative and Procedural Knowledge

Based on the learning materials, tests were developed to assess the declarative and procedural knowledge. The same tests were used before and after learning. Participants were given an "I don't know" option for every question in order to reduce guessing. Declarative knowledge test included 16 questions including 7 multiple choice questions (e.g., What are the advantages of dynamic websites?) each having 4 answer options, and 9 declarative open-ended questions (e.g., Which command is used to define a variable?). Each correctly answered option / question of the declarative questions was rated as one point, leading to a maximum total score of 37. Internal consistency of the declarative knowledge scale was calculated as an indicator for reliability for the pre (Cronbach's α = .92) and post test data (Cronbach's α = .81). Mean test difficulty for declarative knowledge, computed as the ratio of achieved score and maximum score, was .58.

Procedural knowledge test included 4 near transfer questions (e.g., "Write down a JavaScript function that calculates the mean of two given numbers"). For the item construction of near transfer, the description as being "similar to those presented in the booklet and require applying" given by Mayer (1975, p. 531) was followed. A scheme was developed to rate the open-ended questions. As the domain of programming is well-structured and answers were either right or wrong, it was not necessary to have multiple raters. For the transfer questions, each sub-goal was rated as one point (e.g., one point for correct syntax of the programming code and one point for the correct use of a formula) and a maximum total score of 16 could be achieved. Internal consistency of the scale was calculated as an indicator for reliability for the pre (Cronbach's α = .78) and post test data (Cronbach's α = .63). Mean test difficulty for procedural knowledge, computed as the ratio of achieved score and maximum score, was .37.

Evaluation of the Dashboard

After learning, participants in the dashboard and prompt+dashboard groups were presented with a self-report questionnaire to evaluate how useful the dashboard was for their learning process. The questionnaire contained eight 5-step Likert-scaled ("strongly disagree" to "strongly agree") items (e.g., "I adapted my learning behavior due to the information presented in the learning dashboard"). Three items were recoded as they were phrased negatively (e.g., "The dashboard distracted me from learning"). Internal consistency of the scale was calculated as an indicator for reliability (Cronbach's α = .68). Moreover, an open-ended question has been added that asked which parts of the dashboard were helpful or not helpful, reasons for its usefulness, and changes the learners would like to apply. Given answers were clustered manually by their content for an explorative analysis.

Evaluation of the Learning Environment

We presented a questionnaire for the acceptance and technical usability of the learning environment to the control group and the prompt-only group as an alternative task for the evaluation of the dashboard, which was filled by the dashboard groups. The questionnaire included nine 5-step Likert-scaled ("strongly disagree" to "strongly agree") items (e.g., "Using the learning environment was easy"). Four items were recoded as they were phrased negatively (e.g., "I would need technical support to use this learning environment"). Internal consistency of the scale was calculated as an indicator for reliability (Cronbach's α = .88). As for the dashboard evaluation, an open-ended question was included that asked for things the learners would like to change in environment –given answers were clustered for an explorative analysis.

Self-report on Cognitive Load Induced by Dashboard

Cognitive load that learners experienced while using the learning dashboard (dashboard and prompt+dashboard group) was measured with a self-report scale adapted from a preliminary version of the naïve rating questionnaire proposed by Klepsch, Schmitz & Seufert (2017). The scale was designed to measure intrinsic, extraneous, and germane load independently. In the adaption of the scale, ICL was measured with two items, ECL with three items and GCL with two items on a 5-step Likert scale (“strongly disagree” to “strongly agree”). The used items are shown in Table 28. In this study, the internal consistency for all items was Cronbach’s α = .57. Klepsch and colleagues (2017) reported an α value of .86 for their scale, but the lower value could be due to adaptions and an earlier version of the scale that was used. Internal consistency for the subscale ECL was Cronbach’s α = .55 using items ECL-2 and ECL-3, and Cronbach’s α = .46 for GCL using all GCL items. For ICL, only item ICL-1 was used as the two subscale items were correlated negatively. Beside the three subscales, one item was added that directly asked for fun with the dashboard as acceptance of the dashboard was of interest.

Table 28

Items used to measure germane, intrinsic and extraneous cognitive load.

Type / Number	German	English Translation
GCL-1	Beim Durchsehen der Informationen war ich mental angestrengt.	I was mentally strained looking through the information.
GCL-2	Es ging mir beim Durchsehen der Informationen darum, alles richtig zu verstehen.	When I looked through the information, I wanted to understand everything correctly.
ICL-1	Ich musste viele Informationen gleichzeitig im Kopf behalten.	I had to keep a lot of information in my mind at the same time.
ICL-2	Die Informationen zu nutzen war eine sehr komplexe Aufgabe.	Using the information was a very complex task.
ECL-1	Ich habe mich angestrengt, nicht nur einzelne Informationen anzusehen, sondern auch den Gesamtzusammenhang zu verstehen.	I made an effort to not only to process individual pieces of information, but understand the overall context.
ECL-2	Die Darstellung der Informationen ist ungünstig, um mein Lernen nachzuvollziehen.	The presentation of the information is unsuitable to comprehend my learning process.
ECL-3	Es war schwer, die zentralen Informationen miteinander in Verbindung zu bringen.	It was difficult to connect central information with each other.
Fun	Das Durchsehen der Informationen hat mir Spaß gemacht.	Looking through the information was fun.

Evaluation of the Pedagogical Agent

As an alternative to the measurement of CL that participants filled who used the dashboard, participants of the control and prompt-only group were to evaluate the usefulness for the learning process (e.g., “Anna helped me to elaborate the contents”) and the acceptance (e.g., “I liked Anna as my learning assistant.”) of the pedagogical agent with a self-constructed questionnaire including six 5-step Likert-scaled (“strongly disagree” to “strongly agree”) items. Internal consistency of the scale was calculated as an indicator for reliability (Cronbach’s α = .56).

Table 29

Descriptives statistics and Cronbach’s Alpha for the used instruments.

Instrument	N	Items	Min	Max	M	SD	Cronbach’s Alpha
Need for privacy	138	7	1.43	4.43	3.10	.63	.60
Metacognitive Strategies	138	12	2.33	4.58	3.67	.36	.46
Evaluation of Dashboard	65	8	1.63	4.25	3.07	.59	.68
Evaluation of Learning Environment	73	9	1.44	5.00	4.03	.67	.88
Self-Report on Cognitive Load (Mean)	65	7	1.71	4.71	3.27	.59	.57
Germane		2	1	5	3.37	.95	.46
Intrinsic		1	1	5	3.06	1.07	-
Extrinsic		3	1.67	4.67	3.25	.73	.40
Fun-Item		1	1	5	3.26	1.04	-
Evaluation of Pedagogical Agent	73	6	2.17	4.83	3.51	.55	.56
Prior Knowledge							
Declarative	138	16	0	32	5.19	6.19	.92
Procedural	138	4	0	13	.96	2.26	.78
Post Knowledge							
Declarative	138	16	3	36	20.91	5.99	.81
Procedural	138	4	0	16	5.84	4.13	.63

4.5.3 Results

4.5.3.1 Preliminary Analysis

The Type I error rate was set to .05 for all analyses. IBM Statistics 25, PHP, Python, Microsoft Excel and R were used to extract, filter, aggregate and analyze the data set.

Need for data privacy protection and metacognitive strategies will later be used as covariates in hypothesis checks regarding group effects on the declarative and procedural learning outcomes. Thus, these covariates should be independent from a potential group effect. This was checked by computing a one-way MANOVA with group condition as between-subjects factor, and need for privacy, and metacognitive strategies were included as dependent variables. As the equality of covariances is a requirement for the MANOVA, Box's M-test was used which revealed significant violations with Box's $M = 29.13$, $p < .05$. As Tabachnick & Fidell (2013, p. 294) suggest for this case, Pillai's Trace was used as an indicator of significance for the MANOVA, which they described to be robust against this violation. It revealed no significant differences between the four groups regarding the need for privacy and metacognitive strategies, $V = .02$, $F\ (3,133) = 0.35$, $p = .908$. Both learner characteristics variables, need for privacy and metacognitive strategies, are distributed normally according to a non-significant result of the Kolmogorov-Smirnov test and visual inspections of the histograms.

Prior to inferential tests, descriptive statistics and distribution of declarative and procedural knowledge were checked before and after learning. Declarative and procedural knowledge gain were tested for normality by visual inspection and Shapiro-Wilks tests. The tests did provide evidence for normality only for the declarative knowledge gain, but not for procedural knowledge gain. Hence, non-parametric tests will be used for testing group differences in procedural knowledge gain.

Item values of the dashboard evaluation and the CL scale that were presented to the dashboard and prompt+dashboard groups were summed up and descriptive statistics were computed. For the control and the prompt group, item values for the evaluation of the pedagogical agent and the learning environment were summed up and descriptive statistics were computed.

4.5.3.2 Hypothesis Testing

In order to test the two hypotheses *Prompts-And-Dashboard* and *Prompts-Or-Dashboard*, ANCOVAs were first conducted to determine statistically significant differences between all groups regarding declarative learning outcome controlling for the covariates metacognitive strategies and need for privacy. Levene's test of the assumption of homogeneity of variance was evaluated for declarative learning outcome and non-significant results indicated there were no violations of assumptions for ANCOVA. The covariate metacognitive strategies was not significant in the model. Need for privacy was significantly related to the declarative learning outcome, but with a small effect in the model, $F(1,132) = 5.461$, $p < .05$, partial $eta^2 = .040$. The ANCOVA showed no significant differences on declarative learning outcome between the groups, $F(3,132) = .423$, $p = .737$, partial $eta^2 = .010$.

A non-parametric implementation of ANCOVA (Young & Bowman, 1995; implemented in the R package *sm* by Bowman & Azzalini, 2014) was used to check for significant differences between all groups regarding the procedural learning outcome. It revealed no significant differences between the groups, $F\ (3,132) = .262$, $p = .853$, partial $eta^2 = .006$, and no significant effects of the covariates metacognitive strategies and privacy. Figure 35 shows declarative and procedural knowledge before and after learning for each of the four groups. Table 30 shows means and standard deviations for learner characteristics by group.

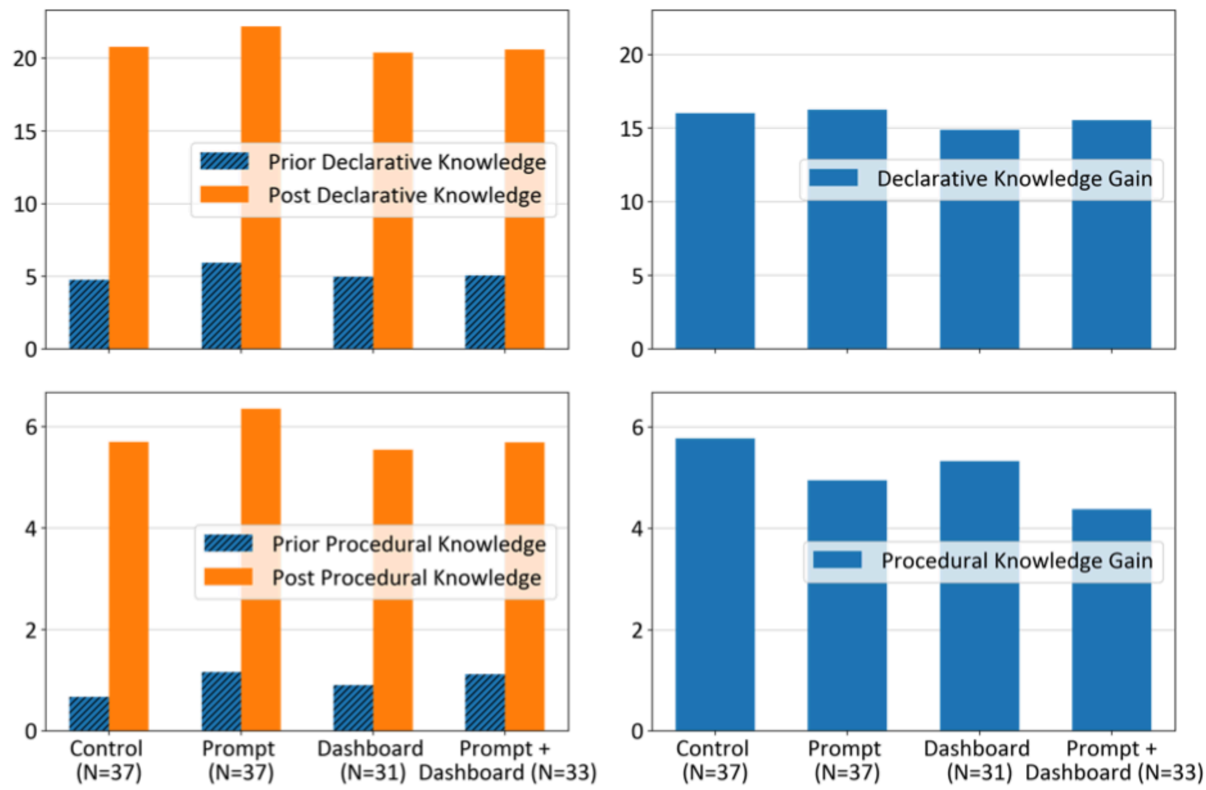


Figure 35. Prior, post and gain of procedural and declarative knowledge for different intervention groups.

Table 30

Declarative and Procedural Knowledge Prior and After Learning by Groups

	Control (N=37)		Prompt (N=37)		Dashboard (N=31)		Prompt + Dashboard (N=33)	
	M	SD	M	SD	M	SD	M	SD
Pre								
Declarative	4.76	7.11	5.95	5.46	4.97	5.21	5.06	6.88
Procedural	0.68	1.45	1.16	2.13	0.90	2.40	1.12	2.96
Post								
Declarative	20.78	5.61	22.19	5.18	20.39	7.55	20.61	6.04
Procedural	5.70	4.45	6.35	4.08	5.55	4.13	5.70	3.99
Gain								
Declarative	16.03	6.80	16.24	6.84	14.90	6.08	15.55	5.96
Procedural	5.03	3.85	5.19	4.03	4.65	3.64	4.58	3.25

Note. Maximum score for declarative knowledge was 37 points, and 16 points for procedural knowledge.

Table 31

Descriptives of learner characteristics by group.

	Control (N=37)		Prompt (N=37)		Dashboard (N=31)		Prompt+Dashboard (N=33)	
	M	SD	M	SD	M	SD	M	SD
Metacognitive Strategies	3.64	0.46	3.68	0.39	3.69	0.32	3.68	0.23
Need for privacy	3.18	0.49	3.12	0.64	3	0.74	3.10	0.66

Aptitude-treatment effects (ATI) regarding the prior knowledge of learners were checked. To do this, a median split was done for declarative and procedural prior knowledge. Values at the median were assigned to the group left of the median. In further analyses, persons with values lower than the median for both variables were treated as having low prior knowledge, whereas persons with at least one variable above the median were not. The frequencies of the resulting groups are shown in Table 32. Mean values for declarative and procedural knowledge gain are shown separately for learners with prior knowledge lower or equal the median, and for those with prior knowledge higher than the median in Figure 36. For the sub-samples with low prior knowledge as well as the remaining learners, differences between the interventions with regards to declarative and procedural learning gains were calculated. There was a significant difference between the intervention groups regarding declarative knowledge gains for learners with low prior knowledge, $F(3, 61) = 3.314$, $p < .05$, $eta^2 = .13$, but not for learners with prior knowledge higher than the median, $F(3,65) = .237$, $p = .863$, $eta^2 = .011$. For learners with little prior knowledge, the contrast regarding declarative knowledge between dashboard and control group is significant and positive (-3.995, $p < .05$), whereas for learners with higher prior knowledge the same contrast is not significant, but the descriptively highest (+1.802, $p = .431$). No other contrast to the control group was significant.

Taking these results into account, the intervention Prompt-and-Dashboard did not improve learning outcomes compared to the other interventions or the control group. Moreover, equipping learners with either a prompt or a dashboard did not improve the learning outcomes compared to no intervention. Hence, both hypothesis, *Prompts-And-Dashboard* and *Prompts-Or-Dashboard* are rejected.

Table 32

Resulting groups of a median split on declarative and procedural knowledge.

		Number of participants	
	Median	< = Median	> Median
Declarative	5	76	62
Procedural	0	99	39
Declarative or Procedural		108	30
Declarative and Procedural		67	71

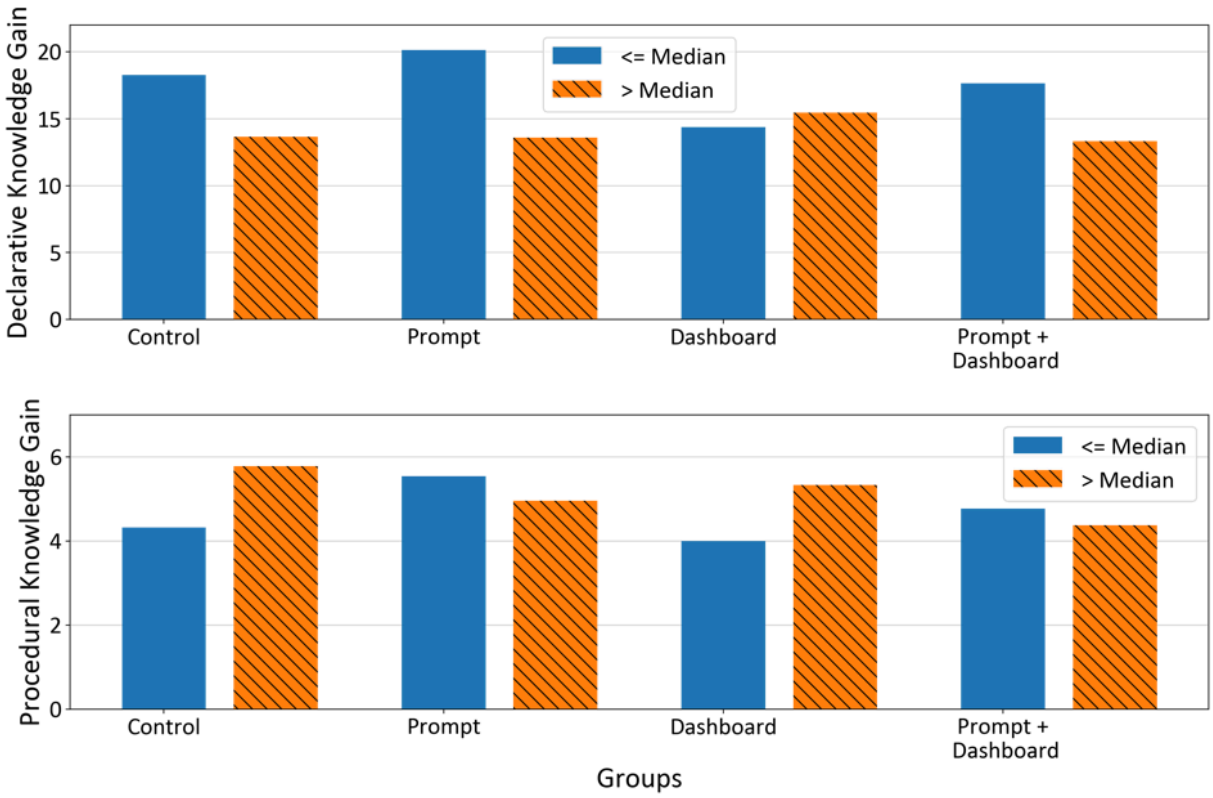


Figure 36. Declarative and procedural knowledge gain for learners with prior knowledge lower or equal the median and for learners with prior knowledge higher than the median.

To check the *More-Non-Linear-Navigation-Hypothesis,* the number of non-linear navigation steps was computed for each learner. Non-linear navigations are deviations from moving to the next page of the given structure of the learning environment, including both back and forward page selections. An ANCOVA was computed to compare the number of non-linear navigation steps between the groups. Non-significant results of Levene's test of the assumption of homogeneity of variance indicated no violation. The covariates metacognitive strategies and need for privacy were not significant in the model. Groups did not significantly differ regarding the steps of non-linear navigations, $F(3,132) = .324$, $p = .808$, *partial eta*^{2} = .007. Hence, the hypothesis is rejected. As shown in Figure 37 groups differ overall on a descriptive level in the order Control < Prompt < Prompt+Dashboard < Dashboard. Moreover, correlations were computed between the number of non-linear navigation events and declarative and procedural knowledge gains. A small significant correlation was found for declarative knowledge gains ($r = .190$, $p < .05$), but not for procedural knowledge gains ($r = .082$, $p = .322$). The number of non-linear navigations steps was also checked for ATI effects. Although not significant, for learners with low prior knowledge, groups differ in the order Control < Prompt < Dashboard < Prompt+Dashboard as expected in the hypothesis, $F(3,61) = .776$, $p = .831$, *partial eta*^{2} =.036. In contrast, for learners with more prior knowledge, the order is inverse, Control > Prompt > Dashboard > Prompt+Dashboard, but also not significant, $F(3,65)= .293$, $p = .512$, *partial eta*^{2} = .013.

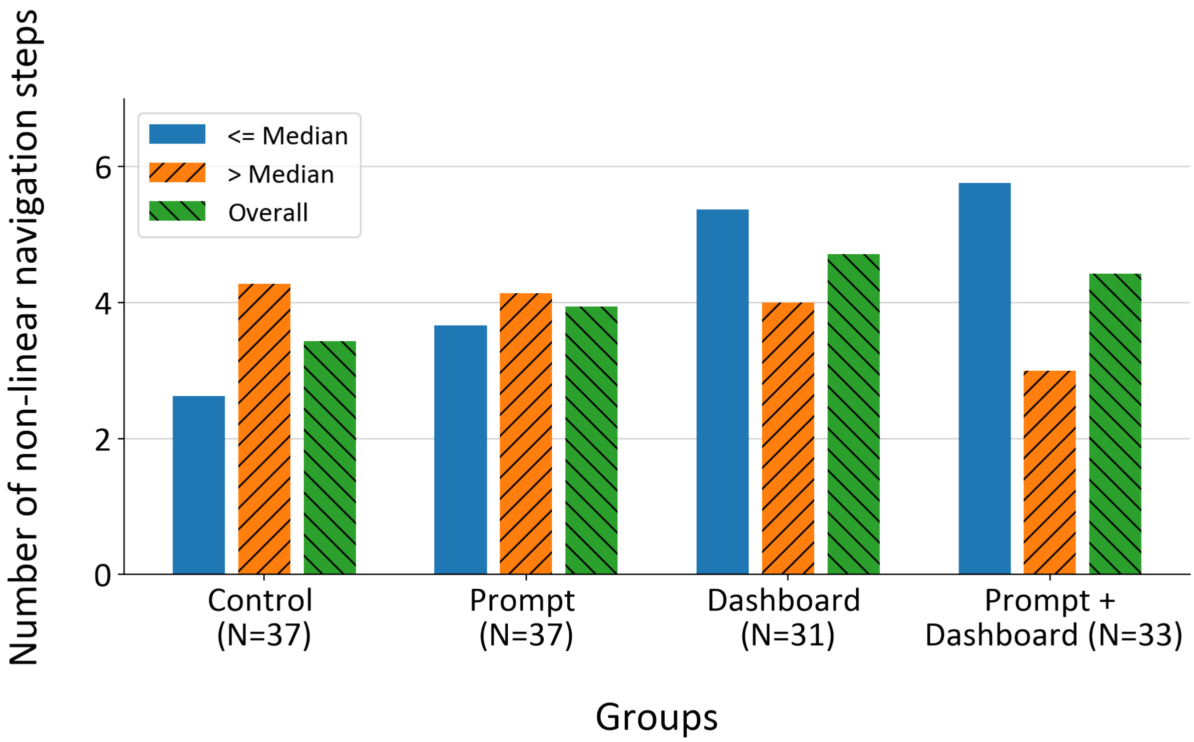


Figure 37. Number of non-linear navigation steps in the learning environment by group, for learners with prior knowledge lower or equal the median, above and overall.

4.5.3.3 Exploratory Analysis

1) How do learners interact with prompts and the learning dashboard?

a. How long and frequently do learners interact with the interventions?

In order to better understand how the prompts were used by the learners, prompt view times were computed as the time difference (in milliseconds) between accessing the dashboard page and closing the prompt window (which occurred automatically after accessing the page). This was done for the prompt group and the prompt+dashboard group. In the prompt group, usage times of 3 prompts occurrences were removed from the analyses as viewing times over 45 seconds without any mouse movements were interpreted as pausing. In the prompt group, due to technical issues, one learner did not receive the first prompt. Four learners did not get the second prompt as they did not reach the page after which the prompt would have been triggered. In the prompt+dashboard group, one learner experienced a usage time of over 45 seconds without showing any interactions and was removed. A total of five learners did not get the second prompt as they did not reach the page after where the prompt would have been triggered. In Table 33, the view times in milliseconds are listed for both prompt occurrences by experimental groups. These are visualized in Figure 38. Differences were significant according to computed paired t-tests, for the group prompt with a medium effect size, $t(29) = 4.260$, $p < .001$, $d = .716$ and for the group prompt+dashboard with a high effect size, $t(26) = 9.729$, $d = 2.148$.

Table 33

View Times of the Prompts by Experimental Groups (in milliseconds)

Group	Prompt Occurrence	N	Min	Max	M	SD
Prompt	First	34	1000	26000	11941	5764
	Second	30	550	16657	3098	3002
Prompt + Dashboard	First	32	3318	32151	15861	7795
	Second	27	907	8240	3601	1873
Overall	First	66	1000	32151	13842	7052
	Second	57	550	16657	3336	2522

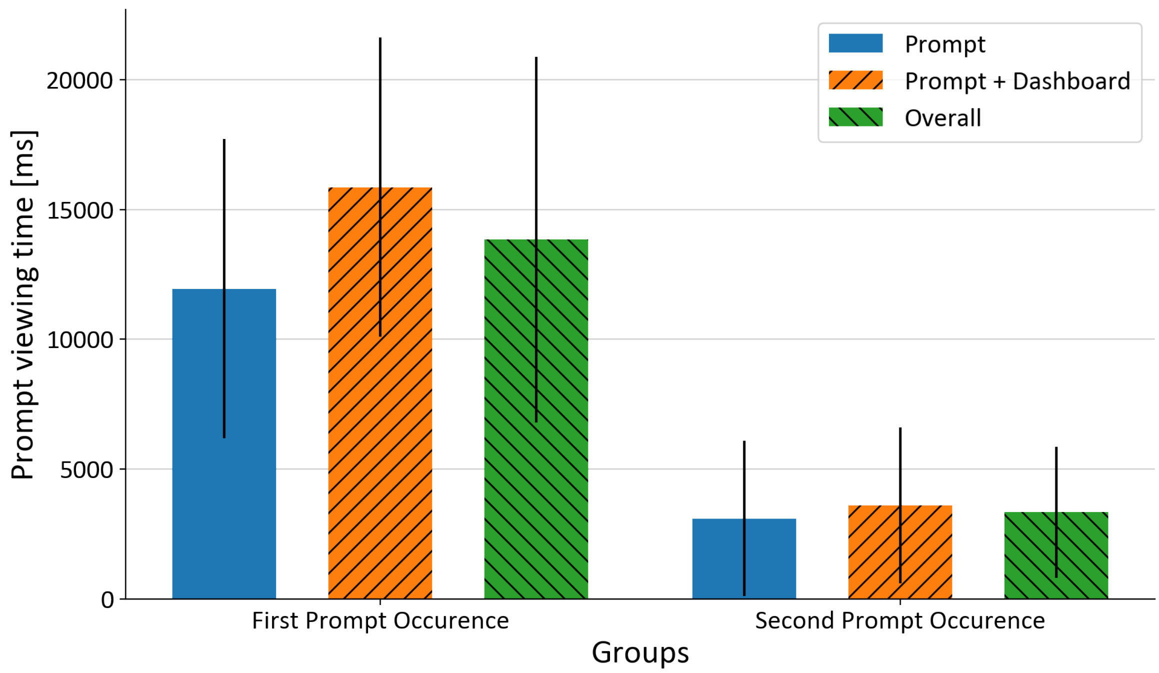


Figure 38. Comparison of first and second occurrence of prompts by groups and overall.

Moreover, Table 34 lists the computed view times of the different tabs in the dashboard that are described above. As the dashboard group did not have prompts, the “questions” tab, where learners could read the prompted questions again, is only listed for the prompt+dashboard group. This is visualized in Figure 39.

Table 34

View Times of Different Tabs in the Dashboard by Experimental Groups

Group	Occurrence	Tab	N	Min	Max	M	SD
Dashboard	First	Information (initial)	31	8514	50487	27569	11314
		Navigation	35	297	43080	15184	11276
		Heat maps	30	713	104550	23954	23311
		Information (revisited)	15	913	24276	7654	7432
Dashboard	Second	Information (initial)	26	254	94665	22254	18922
		Navigation	17	553	50122	10470	11080
		Heat maps	13	105	30522	11335	8561
		Information (revisited)	6	712	10077	4686	3992
Prompt+Dashboard	First	Information (initial)	32	7635	57404	23188	12402
		Navigation	40	200	57493	16192	12572
		Heat maps	39	675	164667	26398	34617
		Questions	39	484	37880	7290	7346
		Information (revisited)	21	115	225188	18286	49276
Prompt+Dashboard	Second	Information (initial)	28	4216	42993	17635	9279
		Navigation	23	924	32973	7996	7866
		Heat maps	24	795	294175	18643	59046
		Questions	17	142	5108	2224	1474
		Information (revisited)	5	760	10195	3547	3846

Note. “Information (initial)” and “Information (revisited)” contain the same content but “Information” was the first visible tab when closing the prompt, so “Information (revisited)” means that learners actively clicked at this tab again, so these were mentioned separately.

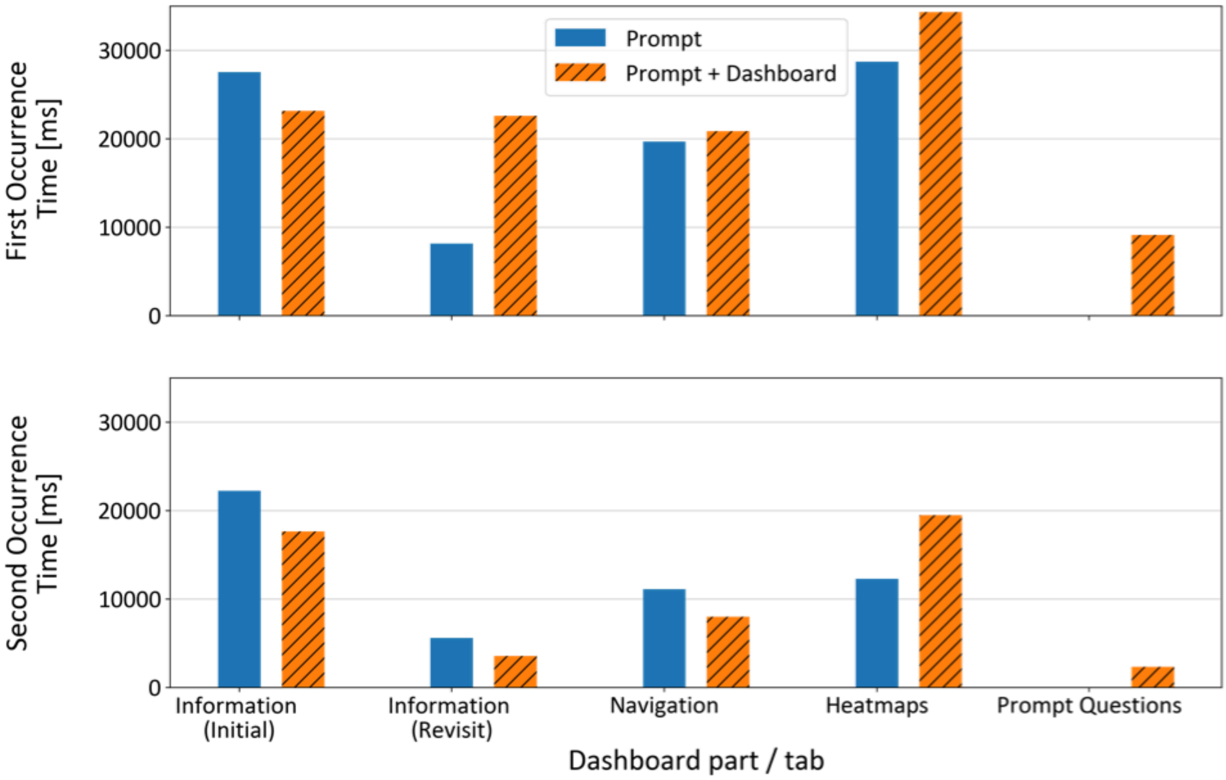


Figure 39. Time spent on different parts of the dashboard for groups dashboard and prompt+dashboard separately for first and second occurrence of the dashboard.

b. What CL is Induced by the Learning Dashboard?

In order to see the level of CL induced by the dashboard, descriptive were computed separately for GCL, ECL and ICL as shown in Table 35. Moreover, all types of CL were compared between the groups dashboard and prompt+dashboard in order to see whether the prompt has an effect on the experienced CL. As shown in Figure 40, learners that received a dashboard without prompts reported lower GCL, ECL, and ICL with small to medium effect sizes, but these differences were not significant.

Table 35

Descriptive Statistics on the Reported Cognitive Load Associated with the Dashboard.

Group (*N*)	Load	*Min*	*Max*	*M*	*SD*
Dashboard	Germane	1	5	3.18	1.05
(*N* = 31)	Extrinsic	1.67	4.67	3.15	.86
	Intrinsic	2	5	3.19	.74
Prompt+Dashboard	Germane	2	5	3.56	.85
(*N* = 33)	Extrinsic	2	4.33	3.34	.59
	Intrinsic	2	4.50	3.23	.69

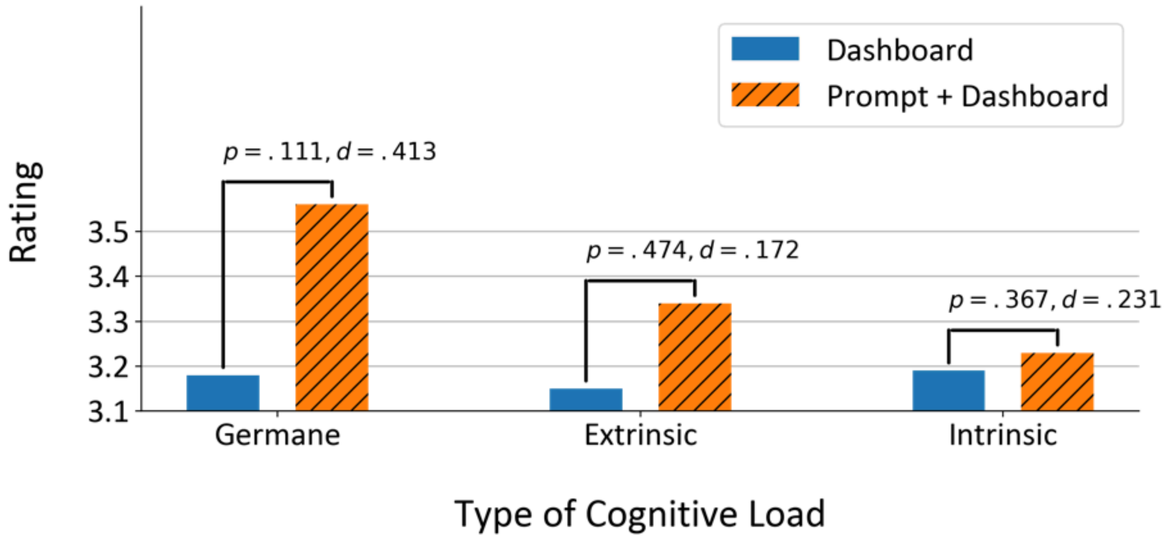


Figure 40. Experienced germane, extraneous, and intrinsic cognitive load separately for dashboard and prompt+dashboard groups.

2) How Do Learners Perceive the Learning Dashboard?

a. Do Learners Perceive the Dashboard as Useful for their Learning?

In order to get insight into how learners perceived the dashboard, descriptives of the dashboard evaluation questionnaire were analyzed on item level and overall, as shown in Figure 41 and Table 36. Mean values for the evaluation of the dashboard did not differ between the groups dashboard and prompt+dashboard. Evaluation of the pedagogical agent and the learning environment presented in Table 36 were not of direct interest for the research questions of this study, but were used as a fill-in for the measures that were presented to the dashboard and prompt+dashboard group.

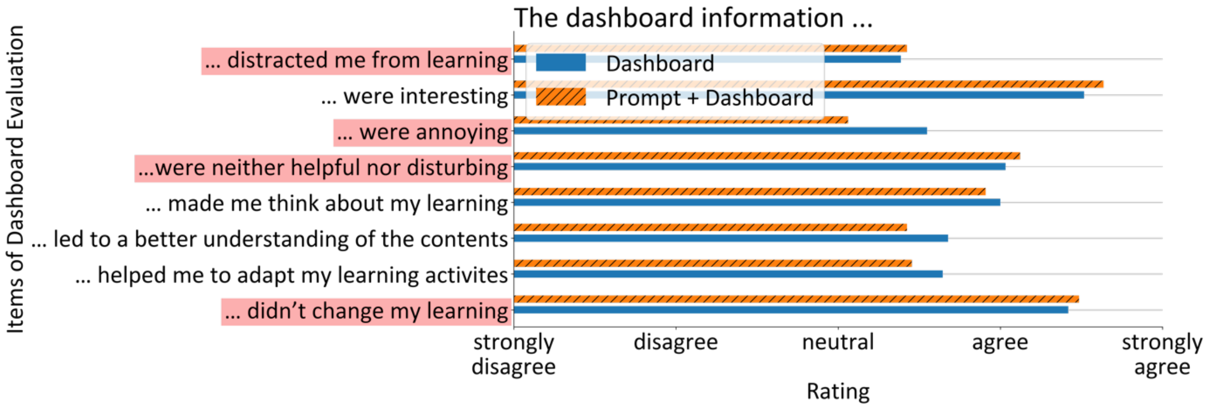


Figure 41. Items of the evaluation of the dashboard for dashboard and prompt+dashboard groups, ordered by the degree of agreement. Labels of negative items are marked with a (red) background and not yet recoded, positive items have no background.

Table 36

Descriptive Statistics on the Evaluation of the Dashboard, the Pedagogical Agent and the Learning Environment.

Questionnaire	Group	N	Min	Max	M	SD
Evaluation Dashboard	Dashboard	32	1.75	4.25	3.07	0.58
	Prompt+Dashboard	33	1.63	4.00	3.07	0.60
Evaluation Pedagogical Agent	Control	37	2.17	4.83	3.56	0.56
	Prompt	37	2.50	4.50	3.45	0.54
Evaluation Learning Environment	Control	37	1.44	4.78	3.98	0.69
	Prompt	37	2.67	5.00	4.08	0.67

b. Which Parts Do Learners Perceive as Useful, and Which Not?

Open answers given in the dashboard evaluation questionnaire were analyzed in order to get more detailed information of what parts of the dashboard were perceived as useful / not useful. To do this, given answers were first categorized through content analysis. A new category was created when statements that did not fit into an existing category were mentioned at least twice. This led to the following five categories: usefulness of information on task progress, usefulness of graphics on navigation, usefulness of heat maps, overall usefulness of the dashboard, usefulness of a comparison with others. These analyses are summed up in Table 37 and visualized in Figure 42. Although there was no actual social comparison built into the dashboard, suggested values for task were perceived as a social comparison by at least 8 learners who mentioned that specifically in their open answers and a category "Comparison with others / with suggested values" was added. Mentioned usefulness in the statements was then rated (1=useful, 0=interesting but not useful, -1=not useful) and counted. A mean index of usefulness was computed as the sum of number of mentions multiplied by the rated usefulness. Heat maps were the most often mentioned part in the dashboard with the lowest usefulness index, and the highest rating to be "interesting". Navigation graphics were mentioned second most often and most positive according to the usefulness index, followed by information on the task progress and the comparison with others which was actually a comparison to suggested values that learners should reach in tasks.

Table 37

Qualitative Analysis of Mentioned Usefulness of Different Dashboard Contents.

Category	Useful (1)	Interesting but not useful (0)	Not useful (-1)	Number of mentions	Usefulness Index
Information on task progress	13	3	3	19	10
Navigation graphics	19	6	7	32	12
Heat maps	4	7	28	39	-24
Dashboard overall	4	0	4	8	0
Comparison with others / with suggested values	10	2	2	14	8

Note. "Useful" was coded as 1, "interesting but not useful" as 0 and "not useful" as -1.

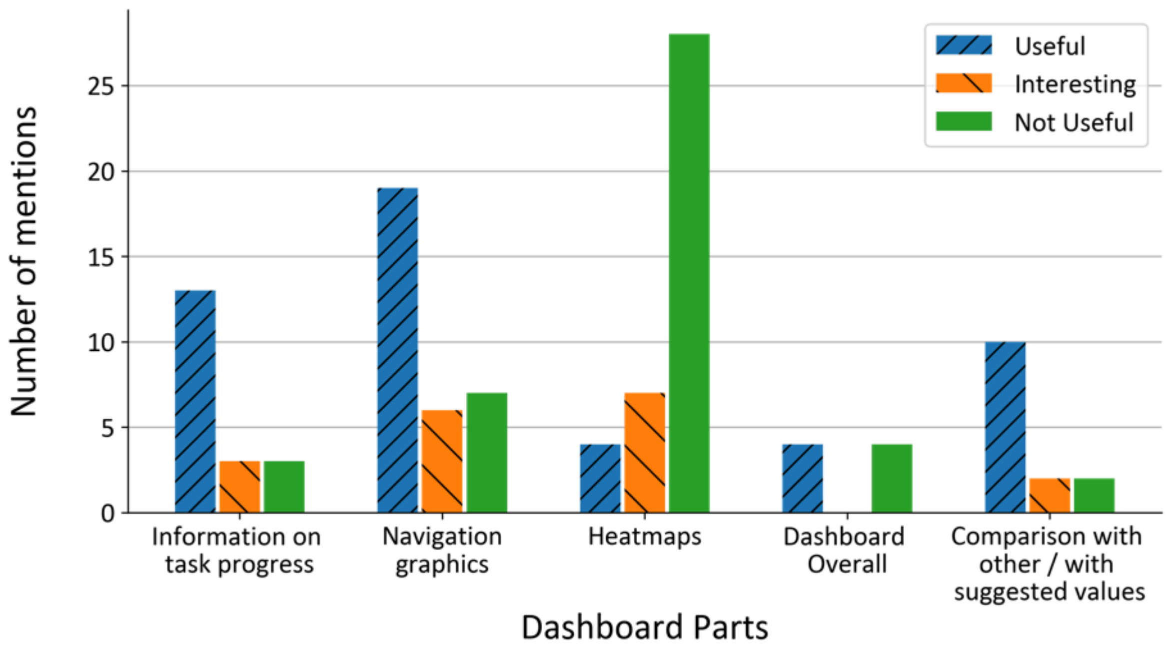


Figure 42. Mentioned usefulness of different dashboard parts in open answers.

4.5.4 Discussion

This study examined the impact of three different interventions, namely prompting, learning dashboards and a combination of both on learning outcomes in an online learning environment on programming. The hypotheses that all three interventions increase the declarative and procedural learning success compared to a control group without intervention (Prompts-Or-Dashboard-Hypothesis) and that a combination of dashboard and prompt has the highest positive effect (Prompts-And-Dashboard-Hypothesis) could not be confirmed. The central concern of this discussion is therefore to uncover possible reasons for these missing effects.

Requirements for Regulation Through Learners

It must first be considered how the interventions should work from the theoretical perspective of SRL in the COPES model (Winne & Hadwin, 1998). Dashboards should assist learners in monitoring their learning process through external information about it. Prompts should solve the production deficit and strengthen the response to changed information. A prerequisite for the effect is therefore the necessity of regulation. If no regulation is necessary, no opportunities are given to apply regulatory activity and hence, the intervention cannot show its effects (e.g., Hadwin et al., 2017). It is suspected that the learning environment has not made enough demands on the regulation of learners. The contents of the learning environment may already have offered a too high level of structure, so that the regulation by the learner himself was not necessary or a lack of regulation was at least not harmful for the learning outcomes. Part of this structure was the linear dependency of the content that arises from the characteristics of the domain of programming. With purely declarative knowledge, such as historical facts, there is little dependency between different knowledge entities - the entities can be learned independently. For the acquisition of programming skills, however, the acquisition of declarative and procedural knowledge is necessary, whose entities each have a high dependency on each other. This is comparable to the concept of element interactivity, which is responsible for high intrinsic loads in the Cognitive Load Theory (Chandler & Sweller, 1996). Since this dependency was considered in the instructional design of the learning environment, it makes sense for the learner to follow the given order of pages. Typical challenging activities of the SRL, such as planning the next learning step and searching for relevant content, could have been greatly facilitated by this structure.

Non-linear Navigation

The results of the *More-Non-Linear-Navigation-Hypothesis* are also relevant for this argumentation. In this hypothesis, learners with intervention were expected to show more non-linear navigation steps, as this indicates systematic learning behavior (Astleitner, 1997) and has already been empirically confirmed for other established interventions (e.g., Bannert et al., 2015; Pieger & Bannert, 2018). Although descriptive and non-significant, more non-linear navigation steps were found for the intervention groups. However, there was no correlation between the number of non-linear navigation steps and the learning outcome in terms of declarative or procedural knowledge. The hypothesis was therefore rejected as well. Looking at these group differences separately for low and high prior knowledge, it can be seen that the number of non-linear navigation steps for learners with low prior knowledge is conform to the hypothesis (Control < Prompt < Dashboard < Prompt+Dashboard). For learners with higher prior knowledge, the order is the exact opposite. This could indicate an ATI effect of the intervention (Snow, 1989), which is reflected in the navigation behavior but not in the learning outcomes. For learners with high prior knowledge, prompt and dashboard may not play an important role, whereas poor learners may benefit from more support from prompt, dashboard, or a combination of both. The fact that the intervention had an effect on navigation behavior but not on learning outcomes may have several reasons. Initially, the differences in the number of non-linear navigation steps were not significant and were accordingly small. On the other hand, it may be that learners were able to make correct metacognitive assessments of their learning due to the interventions, but that the respective contents could still not be learned correctly during re-learning. The explanatory power of non-linear navigation is limited to the fact that learners want to look up the corresponding content again on the basis of a metacognitive evaluation of their lack of understanding. On the other hand, no statement can be made about the quality of the renewed reception of the content. This could also be used to interpret the lack of correlation between the number of non-linear steps and learning outcomes. In addition, this suggests that non-linear navigation was not decisive for high learning outcome due to the linear structure of the learning environment. A comparison of the percentage of non-linear navigation steps also confirms that the content has a high degree of linearity. In this study, this proportion was an average of 4 out of 16 pages, i.e., 25%, whereas in a structurally comparable study using prompting, it was between 51% and 61% (Pieger & Bannert, 2018, p. 170).

Regulation through a pedagogical agent

Another factor that could have limited the need for regulation is the use of an pedagogical agent that gave feedback on the tasks performed. Although it was cognitive feedback on the programming tasks and the dashboard intervention aimed at metacognitive support, the agent may have contributed to facilitating monitoring of the current learning status. Learners were already given feedback on their responses during task completion, so that cognitive evaluations of the achievement of learning objectives may not have been necessary. One indication of this is that feedback from the agent of learners in the control group and the prompt group is moderately more positive than feedback from the dashboard of the corresponding groups. Similarly, other studies find it difficult to disentangle and isolate the effects of cognitive and metacognitive support (Azevedo et al., 2016).

How did learners use prompting?

An important part of this study were the explorative questions. With the data collected on the interaction with the interventions, unresolved questions in this area of research could be answered. As demanded in research on prompting (Bannert & Mengelkamp, 2013), this study investigated how learners used prompts. Firstly, it seems very clear that prompts are only considered by the learner when they first appear. How long a prompt was open differed dramatically between the first and second presentation of the prompt. The first prompt was on average three (group prompt) to five (group prompt+dashboard) times longer open than the second. Far less clear is the interpretation of these different times. Based on this data, no statement can be made about the effect of the second prompt. The learner may need more time to read the prompt when it first occurs. At the next occurrence, the prompt could still have the same effect on the learner. The content of the prompt may no longer be actively processed. Instead, the prompt content is already symbolically represented. It therefore only serves as a trigger for a strategy that has already been acquired. A far less optimistic interpretation, which is just as admissible on the basis of the data, is that the learners simply clicked away the second prompt without benefiting from it in any way.

How Did Learners Use the Dashboard?

The analyses of the process data for interactions are also very relevant for the dashboards. So far, there have been hardly any studies on how dashboards are used. Although the differences are not as extreme as with prompts, there is also a considerable difference between the first and second presentations for dashboards. The second presentation of the dashboard takes far less time than the first.

The dashboard included information on the current status of the tasks, an overview of the navigation behavior, heat maps of the mouse behavior on the pages of the learning environment, and the possibility to re-read the prompt texts for the prompt+dashboard group. The proportional allocation of the time to the different areas of the dashboard is almost the same for both presentations of the dashboards. There is only a small increase for the information on the current status of the tasks and an according reduction for the heat maps. However, the long time that learners spent studying the heat maps during the first presentation is noticeable. This can be interpreted as an indication that learners are not yet familiar with this type of visualization or information and that it therefore has a high salience. On the other hand, the high perception of heat maps as "useless" will be discussed later in the evaluation.

Cognitive Load of Learning Dashboard

Regarding the CL it was found that learners reported higher values for additionally presenting the prompt in the dashboard, but these values were not significantly higher than those of learners without prompt. The largest difference occurs for germane load with medium effect size, the smallest effect size occurs for extraneous load. Since the items' wording was explicitly focused on the load through the dashboard, it is reasonable to assume that the additional presentation of a prompt improves the cognitive processing of the information in the dashboard. There is no evidence for a cognitive overload of learners through the dashboard in the available data. This could also have been a reason for the missing effect of the intervention.

Perceived Usefulness of the Dashboard

In addition to the CL, the perceived usefulness of the information in the dashboard was investigated. A questionnaire was used for the overall dashboard, a specific evaluation of individual components of the dashboard could be reported by the learners in an open question. In the questionnaire there was no mean difference between the prompt group and the prompt+dashboard group. Thus, contrary to CL, the prompts had no influence on the dashboard evaluation. An analysis of the individual items shows that learners found the information in the dashboard interesting, but did not think that it had changed their learning. On the item level, it is also interesting that the group presented with a prompt in addition to the dashboard rated the dashboard information as slightly less annoying. Perceiving the information in the dashboard as annoying may have led to low acceptance of the intervention and is a possible reason for the reduced time at the second presentation of the dashboard.

The open question of the evaluation of the dashboard allows an estimate of which parts the learners found useful. Heat maps were considered to be the least useful. It is unclear, however, whether learners were too challenged with interpreting the visualization, since heat maps are less common than, for example, bar or pie charts, where learners are reported to already experience comprehension problems (e.g., Park & Jo, 2015). The evaluation of the heat maps as interesting speaks for this assumption, the reported values of the CL rather against it. This raises the general question of the meaningfulness of the data channels presented. Although, it was described from a theoretical perspective how the visualizations should act, the effect of each individual data channel visualizations has not yet been empirically clarified and needs further investigation (e.g., Bojko, 2009 for heat maps).

Conclusion and Future Research

The aim of this study was to examine the effects of prompting, dashboards and a combination of both from an SRL perspective, with particular attention given to how interventions are used. No advantages of the intervention compared to the control group could be found. In summary, the following potential reasons for the missing effects of the interventions on learning success could be identified: 1) low demand for regulation due to linear structure of the contents, and due to 2) support from the pedagogical agent, 3) lack of understanding of the heat map visualizations in the dashboard, 4) low rating of usefulness and resulting possible low acceptance for the intervention, 5) decreased time of usage of both interventions in the second presentation. It should also be added that the difficulty of the procedural knowledge test was high. It is possible that effects of the interventions could have been observed more differentiated if the test for procedural knowledge had been easier. The explorative questions in this study could shed light on unanswered questions as to how learners perceive dashboards and, most importantly, how they interact objectively with dashboards and prompts. Important findings were that 1) learners spend considerably more time with the first use of an intervention than with subsequent interventions, 2) high usage time is not necessarily associated with high acceptance, 3) prompts presumably increase the CL used when using dashboards, 4) information on navigation and on the current status of one's own performance in tasks show the highest acceptance values.

Even if the hypotheses could not be confirmed, the data supports at least the structure of the underlying theoretical considerations, namely that dashboards provide external information and prompting leads to a changed reception of this information. For these reasons, however, it could not be shown that the application of the information to regulation can also be changed in the further learning process.

When investigating such interventions, is it possible that the effects depend on the individual prerequisites of learners. Hence, possible ATI effects of the interventions were also considered. Research stresses such effects, taking into account individual differences for a systematically evaluation of treatments (e.g., Snow, 1989). For example, Pieger and Bannert (2018) investigated ATI effects for prompting, and found that learners with less verbal intelligence and reading competence seem to benefit more from metacognitive prompts in online learning environments than students with higher according abilities. Hence, as the intervention in this study also involves prompting, checks for ATI effects were done by comparing the effects of the interventions between learners with very low prior knowledge and those with higher prior knowledge. Although in this study, such effects could only be found regarding the *More-Non-Linear-Navigation-Hypothesis,* it seems to be an important direction to conduct further research for interventions as complex and dynamic as dashboards.

Further research on dashboards is essential. An important, still largely open question in the context of learning dashboards is which visualization in dashboards makes sense for which pedagogical goal. At the moment, findings are not even consistent for an indispensable prerequisite skill of learning dashboards: whether learners are able to correctly interpret commonly used graphs in dashboard visualizations. As examples, Park & Jo (2015) report about learners having problems with graph interpretation while most of the students in the study done by Corrin & Barba (2014) could correctly use the information.

Theoretical work in pedagogical psychology, cognitive sciences and information visualization must be brought together in a meaningful way. On this basis, well-controlled studies on individual visualizations must be carried out in order to provide theoretical- and evidence-based recommendations for the real-world of dashboards. Moreover, researchers and practitioners should be aware of the literature on how well learners perform at interpreting the presented graphs (for a review, see Glazer, 2011) when designing dashboards.

It is also important in this context that the design of dashboard does not ignore fundamental previous results from research on feedback. The wide range of literature on the effects of feedback is also relevant for dashboards, which need to be reconsidered when designing appropriate interventions. The question of alternative ways of presenting data currently visualized in dashboards should also be explored. Data-based wording of individual texts and recommendations that reflect the learning process or stimulate changes in the learning process through intelligent suggestions are possible. This makes sense because researchers and instructional designers are forced to derive relevant suggestions and empirically validated interpretations from the data and not to leave it entirely to the learners to draw their own conclusions. In conclusion, it must be stated that, while there are some promising directions, further empirical evidence is indispensable for the justified use of learning dashboards in productive learning environments.

5 General Discussion

In this work, “peripheral data” was introduced as a data channel that provides detailed, machine-readable information about the interaction between users and websites, or, in the context of technology-enhanced learning and this work, between the learner and the learning environment. This data channel and its’ possibilities have barely been considered by researchers and practitioners in the field of educational psychology in order to contribute to solve the problem and the demand of finding accurate measures that help to understand and promote the mechanisms of SRL. Hence, the aim of this work was to get a better understanding of how peripheral data can be recorded, but also and more importantly, how it relates to variables that are relevant to SRL, and whether it can be used to promote learning by giving learners insight into it. This goal led to three research questions that have been addressed in one development work and four empirical studies. The first question was addressed by the development work, which addressed the theoretical and methodological characteristics of peripheral data and the description of a software and its features:

1) Is peripheral data a suitable data stream to record and analyze the interactions of learners with learning environments?

On this methodological basis, the first three empirical studies (i.e., study 1 in chapter 4.2, study 2 in 4.3, and study 3 in 4.4) addressed the following question by investigating the relation between typing behavior and learning outcomes as well as motivation (study 1), the relation between mouse behavior and CL as well as affective states (study 2), and the possibility to recognize and measure confusion, item difficulty and metacognitive judgements in multi-item scales through mouse behavior (study 3):

2) (How) is peripheral data linked to cognitive, motivational, affective and metacognitive states of learners?

Finally, the last empirical study addressed the impact of the visualization of peripheral data in learning dashboard in combination with metacognitive prompts on the learning outcomes in online learning environments, and thereby addressed the following question:

3) Can learners benefit from presenting them with visualizations of their acquired peripheral data in learning dashboards?

This chapter summarizes and discusses the findings regarding these three questions. First, major findings are summarized and discussed. Secondly, overall methodological considerations and limitations are considered. Finally, conclusions of the work are drawn and possible future directions in this area of research are described.

5.1 Major findings

RQ1: Is peripheral data a suitable data stream to record and analyze the interactions of learners with learning environments?

Regarding the first research question of this study, whether peripheral data is suitable to record and analyze the interaction of learners with learning environments, ScreenAlytics has been developed as a software-framework that implemented the theoretical idea of covering both context and events triggered by the learner through their input devices. The software has been successfully used in every study of this work and features have been added and improved from study to study.

Nevertheless, technological innovations must always reflect critically on whether the enormous effort required for development can be justified in relation to the resulting gain in insight. Does peripheral data really offer the added value that was described theoretically in chapter 4.1.1 compared to screen recordings, simple log files and mouse and keyboard tracking? To answer this question, one should consider how the use of peripheral data has affected the studies in this paper. In the first study, typing behavior was recorded. Unlike classical keystroke logging (e.g., Sullivan & Lindgren, 2006), peripheral data always records the context to which the behavior is related. In this way, the data for capturing the baseline of typing behavior could be easily separated from the data related to the open recall task and the programming tasks. The data from classic keystroke loggers, on the other hand, would have had to be triangulated with the help of screen recordings, so that an alignment would have been possible. In the second study, a measurement of CL in the experimental group was only triggered if the system did not register mouse movements. The implementation of the study would not have been possible without the implementation of peripheral data in this form - since mouse tracking software is usually separated from the learning environment (e.g., Van Waes et al., 2009), dependencies between mouse interaction and reactions in learning environments are very difficult to implement. Peripheral data provides a direct interface to the learning environment, so events can be triggered that depend on the user's interaction and records it simultaneously. Such events are not only useful for experimental designs, but can also be used for interventions in the future. For example, prompts could be triggered depending on certain patterns such as pauses in behavior. A similar approach has already been investigated in the area of online assessment. Prompts were triggered when a learner left the test environment for possible cheating behavior, but it was necessary to program a tool for this purpose (Diedenhofen & Musch, 2016). In the third study it becomes once again very clear what advantage the recording of the context of mouse movements has over classical methods. Again, it would have been necessary to manually triangulate the mouse data with screen recording data in order to determine the time the respondents spent on an element of the questionnaire (e.g., a certain answer option or the question range of an item) - this coding would have required several hours per respondent, which would also have been error-prone and inaccurate. Finally, in the last study, peripheral data was visualized in a learning dashboard. It would not have been possible to design this with classical methods either. The main advantage here is that the captured data is made available to the learning system in real-time for visualizations. Although the interventions of the study did not improve the learning outcome, the presentation of peripheral data in dashboards offers many further application possibilities, which should be systematically empirically tested in further studies.

With regard to the first research question of this thesis, it can be confirmed that peripheral data is well suited for representing the interaction between learners and the learning environment and that the developed software ScreenAlytics is able to record these data and make it available for analyses and interventions.

RQ2: (How) is peripheral data linked to cognitive, motivational, affective and metacognitive states of learners?

It has already been explained why the recording of peripheral data has additional value compared to established traditional methods. Answering the second question about the relationships between peripheral data and relevant variables of the SRL seems to be even more important. On the one hand, because the first question would lose relevance if the recorded data allow insight into behavior, but this behavior does not allow statements about relevant variables. On the other hand, because the third question, whether learners themselves can benefit from the data, would not make sense either. If the recorded data do not contain information on relevant variables for the regulation of learning, it is very unlikely, from a theoretical perspective, that learners can benefit from them for their regulatory activities.

The first study examined the relationship between typing behavior and declarative and procedural learning outcome in acquiring programming skills, i.e., a cognitive aspect of SRL. An adequate measurement of the current learning progress through the behavior of the learner alone, or even a prediction of the later learning outcome, would be extremely helpful for adaptive learning environments (Shute & Zapata-Rivera, 2008). The achieved additional variance explanation of the current learning progress in open recall tasks of up to 23.9%

through the number of pauses and keystrokes is very promising. Furthermore, 28.5% of the variance of procedural learning outcome could be explained by the same indices when writing programming code. Although keystrokes cannot explain the entire variance, the reached R^2 values are considerable and it was not to be expected from a theoretical perspective that the meta information on keystrokes alone could predict the entire learning outcomes. In combination with analyses of the meaning of text, however, this data source can make serious contributions in adaptive learning environments. In addition, an important result of this study is that typing behavior must be interpreted specific to the task, since for the writing of programming code, inverse correlations to the learning progress were shown than for the writing of free text. Unfortunately, there were no comparable studies at the time of documentation to compare the achieved variance explanation.

On the other hand, the first study explored a possible relationship between typing behavior and motivation. Positive correlations between the number of pauses when writing text in the recall task and the initial motivation measurement as well as negative correlations between different indices of the writing of programming code and the current motivation for this task were found. However, the correlations found are small.

In the second study, CL was used to examine cognitive aspects of peripheral data again. Here, the quasi-experimental design of the study seems significant. Not only correlations were found, but it was proven that there is a causal relationship of medium effect size between pauses in the interaction with the learning environment and the CL during this interaction. However, the reasons for this are still unclear and must be clarified in further studies. What exactly happened during the breaks could not be clarified in the study. Although the peripheral data could be used to establish the connection, further triangulations, for example with retrospective or concurrent protocols on thinking aloud, have to be carried out in order to learn more about what happens during the pauses. It is particularly important to find out what type of CL learners have experienced. The question of whether this was productive or unproductive could only be answered to a limited extent with the available data. Compared to previous studies, the relevance of the learning content in the field is particularly noteworthy. Previous research could reveal similar connections only in very controlled laboratory studies with artificial contents (Arshad et al., 2013; Grimes & Valacich, 2015). Moreover, none of the previous studies (Rheem et al., 2018) has tested the actual CL with an existing measurement paradigm such as dual-task. Although theories on CL have been relevant for quite some time, the results of this study appear to be particularly relevant with regard to the recently summarized relationships between SRL and CL (Seufert, 2018), since the model introduced there describes which role the specific CL plays in which phases of SRL.

In addition, the second study also explored the relationship between peripheral data and affective states. It was argued that the intensity, but not the direction of affective states (positive vs. negative) is related to higher indices of mouse movement. For some indices of mouse behavior, correlations with positive and negative affect were demonstrated. It was also confirmed that the correlations do not differ in their direction between positive and negative affect, i.e., as expected, only the intensity but not the direction of affect can be operationalized. However, the results can only be generalized to a limited extent. First, the temporal proximity between the measurement of the affect and the mouse movements was not sufficient. This would be necessary because affective states change quite dynamically as learning progresses (D'Mello & Graesser, 2012). In addition, it is neither empirically nor theoretically clear which indices are relevant for determining the intensity of affect.

In the third study, the focus was on the detection and measurement of confusion as a central epistemic emotion (Pekrun, 2016) by peripheral data. It was especially important to find a suitable setting in which possible correlations can be uncovered and which is still relevant for learning in technology-enhanced environments. Therefore, the interaction with different multi-item scales was investigated. It could be shown that all manipulated items that induced confusion could be recognized only by the deviating interaction, operationalized by different indices of mouse behavior. Furthermore, it could also be shown that the mouse behavior of the participants differs between two levels of confusion. Thus, mouse behavior as an indicator of confusion becomes even more important. Cognitive aspects were also investigated in this study. It was found that higher indices of mouse behavior are associated with higher difficulty. This can be interpreted as longer cognitive engagement with the items. However, the mouse data alone could not predict the absolute objective difficulty of the item. The previously not yet discussed level of metacognition was also investigated in the third study. At this metacognitive level, connections between peripheral data and subjective difficulty assessments of items were investigated. For most of the indices of mouse behavior examined during item response, it was found that these were associated with higher subjective difficulty.

In addition, the mouse data regarding "Feeling-of-Knowing" as dichotomous judgments of whether one knows the answer to a question (yes/no) were examined. Here, a reverse U-shaped relationship between the response time with regard to FOK ratings and the perceived subjective difficulty was assumed. This means that the response times to FOK ratings operationalized by indices of mouse behavior are low if the subjective difficulty is either very high ("This question is very difficult, I have no prior knowledge and therefore know that I do not know the answer") or very low ("This question is very easy, I have enough prior knowledge to judge that I know this answer for sure"). If, on the other hand, learners are not quite sure whether they know the answer because there is relevant prior knowledge, the answer times are higher. This rational was established almost forty years ago (Glucksberg & McCloskey, 1981), but not yet validated by reaction time experiments in online studies. For items that participants rated as more difficult, the data did indeed show longer response times for FOK judgements. For the items assessed as extremely difficult or easy, however, there were no significant correlations between the response times of the FOK judgements and the extreme subjective assessments. This means that, on the basis of this data, the inverse U-shaped correlation described above can only partly be proven.

RQ3: Can learners benefit from presenting them with visualizations of their acquired peripheral data in learning dashboards?

In order to answer the third research question, whether visualizations of learners' peripheral data in learning dashboards can improve their learning outcomes, an intervention study was conducted. As dashboards as interventions for learning in technology-enhanced learning environments came up only recently, there are still several research gaps that have been identified in recent reviews (Bodily & Verbert, 2017; Gaševic et al., 2015; Jivet et al., 2017; Schwendimann et al., 2017). This study accounted for some of them. It tried to build on a clear theoretical foundation both for the overall mechanisms of a learning dashboard using the COPES-model of SRL (Winne & Hadwin, 1998), as well as for the different kind of information and visualizations in the dashboard. Moreover, instead of just presenting the information to raise awareness (as criticized by Jivet et al., 2017), it implemented prompts that were meant to enhance the usage of the information in learning strategies. Maybe the most relevant claim in previous literature on learning dashboards is the lack of systematic experimental research designs on the effects on learning outcomes that incorporate control groups. Thus, the study tested the effects by implementing an experimental field study and systematically varied the factors prompting and learning dashboards among three intervention and a control group. However, the interventions did not show the hypothesized effects, i.e., neither learning dashboards, nor prompt, nor a combination of both could significantly improve learning compared to the control group. Therefore it was important to look for the potential reasons of the missing effects, such as insufficient opportunities for regulation caused by a linear content dependency and too much support through a pedagogical agent. A key contribution of the work was the analysis of the process data. How learners interact with the dashboards has not yet been examined satisfactorily and is demanded in reviews (e.g., Bodily & Verbert, 2017) to create directions for further research. Important findings here were that 1) learners used the intervention intensively, especially during the first occurrence, 2) the interventions did not lead to cognitive overload, and 3) prompts potentially increased the cognitive processing (operationalized as self-reports on CL) of studying the information in dashboards.

The last point is particularly noteworthy. The hypotheses argued that prompts can help to resolve learners' production deficits and thus encourage learners to actually apply strategies. It was assumed that the strategies refer to the actual learning content and not to the information in the dashboard. But if the prompts have affected the processing of the dashboard content, then further questions arise: 1) Have learners not processed the dashboard content sufficiently without prompts? 2) Should prompts be formulated differently or positioned elsewhere to encourage the application of strategies to the actual learning content?

In addition, it is also relevant for future studies which information of the dashboard was used by learners in what way. It was found that although heat maps on mouse behavior were used for a long time, they were not considered as helpful by learners. The presentation of mouse movements is also interesting with regard to EMME research (chapter 2.5.5). Although this is not particularly in the sense of a learning dashboard anymore, it is very interesting, whether heat maps are helpful, if not the own mouse behavior, but the behavior of particularly good learners in the sense of a Mouse-Movement-Modeling-Example (MMME) is displayed. Similar to EMMEs, cognitive processing could be improved either by the social character of a virtual expert that moves his or her mouse (Krebs et al., 2018) or by adopting a new strategy in learning sessions (Mason et al., 2015). It would also be very interesting to see whether the presentation of a video of one's own learning session with mouse positions instead of heat maps would be helpful. The major difference between the two visualizations is that the information is aggregated in heat maps, because the time sequence of mouse positions is ignored. This aggregation of the information was meant to be useful, since parts that were visited frequently with the mouse pointer become visible, but also parts that were not visited at all.

As a further part of the dashboard, information about the current learning status, for example the number of tasks solved, was perceived as helpful and also used for a comparatively long time. It should be noted here that although the number of correct solutions can be recorded using the ScreenAlytics software framework, these are not purely peripheral data, but rather results from algorithms for evaluating the tasks. Finally, the information on the navigation process was also actively used and well rated.

5.2 Methodological Considerations

Samples and Online Acquisition of Participants

Selection bias in samples can seriously compromise the internal validity of empirical studies (Larzelere, Kuhn & Johnson, 2004). As the first two studies of this work were conducted with university students in media communication, gender was not equally distributed. Moreover, the samples in this work were biased regarding their age as participants were acquired mostly among university students in their earlier semesters. Although there is no general gender or age effect known for the used measures and interventions, this limits the generalizability of the results to other populations of learners.

In general, researchers are very skeptical and have healthy reservations towards to use of online experiments. Even advocates of online research claim that “this mode of research has some inherent limitations due to lack of control and observation of conditions” (Reips &

Birnbaum, 2011, p. 563). Although researchers face some important challenges, I argue that online research provides a huge opportunity for psychology. It is only through the use of online research that both empirical field and controlled “internet-lab” studies can be conducted with large, heterogeneous, easily and rapidly accessible samples while still being economically feasible. This is absolutely necessary to restore the reputation of experimental-psychological research after its’ replication crisis. Promising work is currently done to build evidence of the possibility to conduct “online-lab” studies that meet at least the same quality standards as traditional lab studies (e.g., de Leeuw, 2015; Hilbig, 2015; Semmelmann & Weigelt, 2017). However, researchers must find methods to better control the quality of online samples. The development of ScreenAlytics in this work contributed to this by enabling researcher to conduct detailed quality checks on the basis of interaction data as described in chapter 4.1.5.1.

Learning Materials

Materials that learners were asked to study in study 1, 2 and 4 were about website programming. The domain of programming was chosen as it requires both declarative knowledge on syntax, rules, concepts, and procedural knowledge in order to actually write working code. Moreover, the domain is rather well structured, making it possible to automatically provide feedback on code as done in the studies, but also allowing for accurate measurement of the learning outcomes. Recorded code writing also allows to reconstruct steps towards a solution, thus giving insight into cognitive processes during development. However, this domain also has specific demands to cognitive operations of learners (Jones & Burnett, 2008; Mayer, 1981; White & Sivitanides, 2009) and thus, generalization to other domains are not easily possible.

In addition to the content of the learning materials, the structure of the learning environment needs to be considered. In the domain of programming, understanding a knowledge entity usually depends on already knowing other entities. This domain structure means that in a learning environment all content is relevant and a knowledge entity is built on the one previously learned. This may have led to lower regulation requirements in the studies. More independent knowledge entities require learners to apply more metacognitive and regulatory activities, such as judging the relevance of a content element or planning the next step in the learning process. Although this was probably not problematic for studies in which the relation between peripheral data and SRL variables was investigated, it may have limited the results of the intervention study because less opportunities for learners were provided to actually regulate (Hadwin et al., 2017).

The Issue of Circular Reasoning

When trying to develop a new proxy measures for an established measure of a latent psychological construct, it is indispensable that the established measure itself is reliable and valid. As described in chapter 2.4 (and inspired by Reimann et al., 2014), the studies 1, 2 and 3 of this work identified observable behaviors in peripheral data (i.e., indices of mouse and keyboard behavior in relation to its context), and tried to link these to latent variables. These latent variables were measured using established and self-created instruments. The rationale behind this is, that peripheral data can then measure the latent variable in an unobtrusive way and without using the established measures. However, a major question is to which degree the instruments that peripheral data has been aligned to, are reliable and valid. Thus, the instruments of the studies are briefly discussed regarding their quality.

In the first study, the instrument to assess declarative and procedural knowledge prior to and after learning were self-created and not tested on a large sample before. However, the Cronbach’s alpha value for the declarative knowledge test for internal consistency was medium (post-test) to high (pre-test). Regarding the procedural knowledge test, an authentic web design task was chosen that was rated manually. The coding was done by two raters that achieved a high interrater reliability. Regarding motivation, the initial measure was assessed using an established measure (QCM by Rheinberg et al., 2001) that showed high internal consistency in the data of this study. This test was adapted to a short three-item measure to assess the current motivation in close time proximity to the typing behavior and also reached acceptable values for internal consistency.

In the second study, it was important to close the research gap of aligning mouse movements to an objective measure of CL, as other studies only used subjective self-reports. The use of the dual-task paradigm with a secondary reaction-time task has been intensively described as reliable and valid in literature (e.g., Brünken, Plass & Leutner, 2004; Brünken et al., 2003; Schoor et al., 2012). However, as discussed, a major drawback of this approach is that it does not tell whether the measured load is productive or not. Thus, it seems to be very difficult, also from a theoretical perspective, to align mouse behavior to a specific kind of CL. Regarding the affect measure, high values for internal consistency could be reported, but the affect could not be measured in close time proximity.

In the third study, a major concern was how to induce confusion and, even more importantly, how to control for the successful induction. On the one hand, there is evidence that contradictions and (grammar) errors lead to confusion (D'Mello & Graesser, 2014). On the other hand, the pre-test of the manipulation did 1) not have a large enough sample, and 2) the question whether an item is confusing was a very artificial task. Otherwise, the adapted BEFKI test to assess crystalline intelligence and the BFI-2 has been tested on large samples.

To sum up, in psychology, self-reports are often the only available measure and there is no straightforward way to proof the quality of it. Researchers simply need to trust what learners tell them. An argument for using new data channels to measure SRL variables that is often mentioned in the discourse is that self-reports are criticized for being unreliable and subjective. Hence, another question is whether researchers are able find other ways to avoid using the approach of this study trying to find proxy measures for self-reports. An alternative to this approach would be a sound theoretically and empirically grounded reason to trust a specific pattern in data-channels more than a labelled proxy measure of self-reported data.

5.3 Conclusion and Outlook

The measurement of SRL processes in their various phases at cognitive, metacognitive, motivational and affective levels in learning with technology-enhanced environments is currently a major challenge in educational psychology. This measurement is necessary to better understand SRL on the one hand and to better support learners on the other hand. Thus, researchers claim a demand of objective, reliable, and valid process measures of SRL that are, moreover, available in real-time, unobtrusive and non-reactive to other measures (Azevedo, 2015; Azevedo & Greene, 2010; Sonnenberg & Bannert, 2018; Winne & Perry, 2000).

This thesis tried to further close this research gap by introducing and investigating peripheral data as source to measure and support SRL. It contributed to the discourse from three perspectives. First, by developing and evaluating a software framework that allows researchers and practitioners to capture and pre-process the interactions between learners and learning environment. Secondly, it contributed with the investigation of the relationship between peripheral data and learning outcomes, CL, motivation, and affective states in authentic learning environments as well as confusion, experienced subjective difficulty and objective difficulty and metacognitive judgements in surveys. Finally, it contributed to the emerging area of learning dashboards by investigating the effects and the usage of them with a systematic experimental study that has been demanded by several recent reviews.

However, more research is still needed on all three levels. First, further features need to be implemented in ScreenAlytics, especially a browser plugin that allows researchers in the lab to record websites they do not administrate. This is important to examine SRL processes outside the boundaries of a closed learning environment. Secondly, although important insights could be acquired, the relations between peripheral data and all described latent variables need to be understood better. The results of the studies can give directions for further studies, especially regarding which indices of peripheral data are best suited for further investigations of a specific latent variable. Machine learning algorithms that are used more and more in the field of learning analytics are very promising to confirm and extend the results of the studies of this work. Finally, peripheral data sets the stage for many innovative interventions. A few very interesting and promising examples area sophisticated learning dashboards, pedagogical agents with the ability to replay parts of a learning session to actively provide learners with feedback on their screen, or using recorded learning sessions as a model in the sense of EMMEs to foster the acquisition of new learning strategies.

References

Alimadadi, S., Sequeira, S., Mesbah, A. & Pattabiraman, K. (2014). Understanding JavaScript Event-Based Interactions. *Proceedings of the ACM/IEEE International Conference on Software Engineering (ICSE)*, 367–377. doi: 10.1145/2568225.2568268

Allen, L. K., Mills, C., Jacovina, M. E., Crossley, S., D'Mello, S. & McNamara, D. S. (2016). Investigating boredom and engagement during writing using multiple sources of information. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*, 114–123. doi: 10.1145/2883851.2883939

Anjewierden, A., Kollöffel, B. & Hulshof, C. (2007). Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes. *CEUR Workshop Proceedings*, *305*, 23–32. doi: 10.1109/PIMRC.2014.7136286

Antonenko, P., Paas, F., Grabner, R. & van Gog, T. (2010). Using Electroencephalography to Measure Cognitive Load. *Educational Psychology Review*, *22*(4), 425–438. doi: 10.1007/s10648-010-9130-y

Arguel, A., Lockyer, L., Kennedy, G., Lodge, J. M. & Pachman, M. (2018). Seeking optimal confusion: a review on epistemic emotion management in interactive digital learning environments. *Interactive Learning Environments*, *27*(1), 1–11. doi: 10.1080/10494820.2018.1457544

Arguel, A., Lockyer, L., Lipp, O. V., Lodge, J. M. & Kennedy, G. (2017). Inside Out: Detecting Learners' Confusion to Improve Interactive Digital Learning Environments. *Journal of Educational Computing Research*, *55*(4), 526–551. doi: 10.1177/0735633116674732

Arroyo, E., Selker, T. & Wei, W. (2006). Usability tool for analysis of web designs using mouse tracks. In *Extended abstracts on Human factors in computing systems* (pp. 484–489). doi: 10.1145/1125451.1125557

Arshad, S., Wang, Y. & Chen, F. (2013). Analysing mouse activity for cognitive load detection. In *Proceedings of the 25th Australian Computer-Human Interaction Conference on Augmentation, Application, Innovation, Collaboration* (pp. 115–118). doi: 10.1145/2541016.2541083

Astleitner, H. (1997). *Lernen in Informationsnetzen. Theoretische Aspekte und empirische Analysen des Umgangs mit neuen Informationstechnologien aus erziehungswissenschaftlicher Perspektive.* Frankfurt: Europäischer Verlag der Wissenschaften.

Atkinson, R. . C. & Shiffrin, R. M. (1971). The Control Process of Short-Term Memory. *Scientific American*, *225*(2), 82–91. doi: 10.1038/scientificamerican0871-82

Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology*, *94*(2), 416–427. doi: 10.1037//0022-0663.94.2.416

Atterer, R., Wnuk, M. & Schmidt, A. (2006). Knowing the user's every move. In *Proceedings of the 15th international conference on World Wide Web - WWW '06* (pp. 203–213). New York, New York, USA: ACM Press. doi: 10.1145/1135777.1135811

Attig, C., Then, E. & Krems, J. F. (2018). Mausparameter als Indikatoren für Hilfsbedürftigkeit in der MCI. In R. Dachselt & G. Weber (Eds.), *Mensch und Computer 2018 - Tagungsband*. Bonn: Gesellschaft für Informatik e.V. doi: 10.18420/muc2018-mci-0326

Azevedo, R. (2009). Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: A discussion. *Metacognition and Learning*, *4*(1), 87–95. doi: 10.1007/s11409-009-9035-7

Azevedo, R. (2015). Defining and Measuring Engagement and Learning in Science: Conceptual, Theoretical, Methodological, and Analytical Issues. *Educational Psychologist*, *50*(1), 84–94. doi: 10.1080/00461520.2015.1004069

Azevedo, R. & Greene, J. A. (2010). The Measurement of Learners' Self-Regulated Cognitive and Metacognitive Processes While Using Computer-Based Learning Environments. *Educational Psychologist*, *45*(4), 203–209. doi: 10.1080/00461520.2010.515935

Azevedo, R., Johnson, A., Chauncey, A. & Burkett, C. (2010). Self-regulated Learning with MetaTutor: Advancing the Science of Learning with MetaCognitive Tools. In M. S. Khine & I. M. Saleh (Eds.), *New Science of Learning: Computers, Cognition and Collaboration in Education* (pp. 225–247). New York: Springer. doi: 10.1007/978-1-4419-5716-0_11

Azevedo, R., Landis, R. S., Feyzi-Behnagh, R., Duffy, M., Trevors, G., Harley, J. M., …

Hossain, G. (2012). The effectiveness of pedagogical agents' prompting and feedback in facilitating co-adapted learning with MetaTutor. *Lecture Notes in Computer Science*, *7315*, 212–221. doi: 10.1007/978-3-642-30950-2_27

Azevedo, R., Martin, S. A., Taub, M., Mudrick, N. V., Millar, G. C. & Grafsgaard, J. F. (2016). Are Pedagogical Agents' External Regulation Effective in Fostering Learning with Intelligent Tutoring Systems? In A. Micarelli, J. Stamper & K. Panourgia (Eds.), *Intelligent Tutoring Systems. ITS 2016. Lecture Notes in Computer Science.* (Vol. 9684, pp. 197–207). Springer International Publishing. doi: 10.1007/978-3-319-39583-8_19

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Baker, R. S. J. d., D'Mello, S. K., Rodrigo, M. M. T. & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223–241. doi: 10.1016/j.ijhcs.2009.12.003

Baltrusaitis, T., Robinson, P. & Morency, L.-P. (2016). OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision - WACV* (pp. 1–10). IEEE. doi: 10.1109/WACV.2016.7477553

Bandura, A. (1986). *Social foundations of thought and action : a social cognitive theory*. Englewood Cliffs, N.J.: Prentice-Hall.

Bannert, M. (2002). Managing cognitive load - recent trends in cognitive load theory. *Learning and Instruction*, *12*(1), 139–146. doi: 10.1016/S0959-4752(01)00021-4

Bannert, M. (2007). *Metakognition beim Lernen mit Hypermedien [Metacognition in learning with hypermedia. Assessment, description, and mediation of effective metacognitive learning strategies and regulation activities]*. Münster: Waxmann Verlag.

Bannert, M. (2009). Promoting Self-Regulated Learning Through Prompts. *Zeitschrift Für Pädagogische Psychologie*, *23*(2), 91–94. doi: 10.1024/1010-0652.23.2.91

Bannert, M., Hildebrand, M. & Mengelkamp, C. (2009). Effects of a metacognitive support device in learning environments. *Computers in Human Behavior*, *25*(4), 829–835. doi: 10.1016/j.chb.2008.07.002

Bannert, M. & Mengelkamp, C. (2008). Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does the verbalisation method affect learning? *Metacognition and Learning*, *3*(1), 39–58. doi: 10.1007/s11409-007-9009-6

Bannert, M. & Mengelkamp, C. (2013). Scaffolding Hypermedia Learning Through Metacognitive Prompts. In R. Azevedo & V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies* (pp. 171–186). New York: Springer Science+Business Media. doi: 10.1007/978-1-4419-5546-3_12

Bannert, M., Sonnenberg, C., Mengelkamp, C. & Pieger, E. (2015). Short- and long-term effects of students' self-directed metacognitive prompts on navigation behavior and learning performance. *Computers in Human Behavior*, *52*, 293–306. doi: 10.1016/j.chb.2015.05.038

Banse, R. & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636. doi: 10.1037/0022-3514.70.3.614

Bartlett, F. (1932). *Remembering*. Cambridge: Cambridge University Press.

Batliner, A., Steidl, S., Hacker, C. & Nöth, E. (2008). Private emotions versus social interaction: A data-driven approach towards analysing emotion in speech. *User Modeling and User-Adapted Interaction*, *18*(1–2), 175–206. doi: 10.1007/s11257-007-9039-4

Baylor, A. L. & Ryu, J. (2003). The effects of image and animation in enhancing pedagogical agent persona. *Journal of Educational Computing Research*, *28*(4), 373–394.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure ofprocessing resources. *Psychological Bulletin*, *91*(2), 276–292.

Beaudoin, L. P. & Winne, P. (2009). nStudy: An Internet tool to support learning, collaboration and researching learning strategies. In *Canadian e-Learning Conference*. Vancouver, Canada.

Bergadano, F., Gunetti, D. & Picardi, C. (2003). Identity verification through dynamic keystroke analysis. *Intelligent Data Analysis*, *7*, 469–496.

Berthold, K., Nückles, M. & Renkl, A. (2007). Do learning protocols support learning strategies and outcomes? The role of cognitive and metacognitive prompts. *Learning and Instruction*, *17*, 564–577. doi: 10.1016/j.learninstruc.2007.09.007

Bertrand, M. & Mullainathan, S. (2001). Do People Mean What They Say? Implications For Subjective Survey Data. *SSRN Electronic Journal*, *91*(2), 67–72. doi:

10.2139/ssrn.260131

Birnbaum, M. H. (2004). Human Research and Data Collection via the Internet. *Annual Review of Psychology*, *55*(1), 803–832. doi: 10.1146/annurev.psych.55.090902.141601

Bjork, E. E. & Bjork, R. (2011). Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew & J. R. Pomerantz (Eds.), *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society* (pp. 56–64). New York: Worth.

Bodily, R. & Verbert, K. (2017). Review of Research on Student-Facing Learning Analytics Dashboards and Educational Recommender Systems. *IEEE Transactions on Learning Technologies*, *10*(4), 405–418. doi: 10.1109/TLT.2017.2740172

Boekaerts, M. (2007). Self-Regulation and Effort Investment. In E. Sigel & K. . Renninger (Eds.), *Handbook of Child Psychology* (Vol. 4, pp. 345–377). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi: 10.1002/9780470147658.chpsy0409

Boekaerts, M. & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology*, *54*(2), 199–231. doi: 10.1111/j.1464-0597.2005.00205.x

Boekaerts, M. & Niemivirta, M. (2000). Self-Regulated Learning. In M. Boekaerts, P. R. Pintrich & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 417–450). London: Academic Press. doi: 10.1016/B978-012109890-2/50042-1

Bojko, A. (2009). Informative or Misleading? Heatmaps Deconstructed. In J. A. Jacko (Ed.), *Lecture Notes in Computer Science* (Vol. 5610, pp. 30–39). doi: 10.1007/978-3-642-02574-7_4

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M. & MacKenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: a meta-analysis. *Personell Psychology*, (48), 587–605.

Borkowski, J. G. (1996). Metacognition: Theory or chapter heading? *Learning and Individual Differences*, *8*(4), 391–402. doi: 10.1016/S1041-6080(96)90025-4

Bowker, G. C., Brine, K. R., Gruber Garvey, E., Gitelman, L., Steven J. Jackson, Jackson, V., … Williams, T. D. (2013). *"Raw Data" Is an Oxymoron*. London: The MIT Press. doi: 10.1080/1369118X.2014.920042

Bowman, A. W. & Azzalini, A. (2014). R package sm: nonparametric smoothing methods (version 2.2-5.4). University of Glasgow, UK and Universita di Padova, Italia.

Bradley, M. M., Miccoli, L., Escrig, M. A. & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607. doi: 10.1111/j.1469-8986.2008.00654.x

Brünken, R., Plass, J. A. N. L. & Leutner, D. (2004). Assessment of Cognitive Load in Multimedia Learning with Dual-Task Methodology : Auditory Load and Modality Effects. *Instructional Science*, (32), 115–132.

Brünken, R., Plass, J. L. & Leutner, D. (2003). Direct Measurement of Cognitive Load in Multimedia Learning. *Educational Psychologist*, *38*(1), 53–61. doi: 10.1207/S15326985EP3801_7

Brünken, R., Steinbacher, S., Plass, J. L. & Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental Psychology*, *49*(2), 109.

Chandler, P. & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology*, *10*(2), 151–170. doi: 10.1002/(sici)1099-0720(199604)10:2<151::aid-acp380>3.0.co;2-u

Chen, G., Lee, J., Wang, C., Chao, P., Li, L. & Lee, Y. (2012). An empathic avatar in a computer-aided learning program to encourage and persuade learners empathic avatar design. *Educational Technology & Society*, *15*, 62–72.

Chen, M., Anderson, J. & Sohn, M. (2001). What can a mouse cursor tell us more? Correlation of eye/mouse movements on web browsing. *CHI Shorttalks: Extended Abstracts*, 281–282.

Cheng, G. (2017). The impact of online automated feedback on students' reflective journal writing in an EFL course. *Internet and Higher Education*, *34*, 18–27. doi: 10.1016/j.iheduc.2017.04.002

Choi, S. & Clark, R. E. (2006). Cognitive and affective benefits of an animated pedagogical agent for learning English as a second language. *Journal of Educational Computing Research*, *34*(4), 441–466.

Clarebout, G., Elen, J. & Johnson, W. (2002). Animated pedagogical agents: where do we stand? In *World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 2–8).

Clark, R. E. & Choi, S. (2007). The questionable benefits of pedagogical agents: Response to Veletsianos. *Journal of Educational Computing Research*, *36*(4), 379–381.

Cocea, M. & Weibelzahl, S. (2006). Can log files analysis estimate learners' level of motivation? In K. D. Althoff & M. Schaaf (Eds.), *Lernen-Wissensentdeckung - Adaptivität 2006* (pp. 32–35). Hildesheim.

Cocea, M. & Weibelzahl, S. (2007a). Cross-system validation of engagement prediction from log files. *Proceedings of Second European Conference on Technology Enhanced Learning*, 14–25. doi: 10.1007/978-3-540-75195-3_2

Cocea, M. & Weibelzahl, S. (2007b). Eliciting Motivation Knowledge from Log Files Towards Motivation Diagnosis for Adaptive Systems. In C. Conati, K. McCoy & G. Paliouras (Eds.), *User Modeling 2007* (Vol. 4511, pp. 197–206). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-540-73078-1_23

Cocea, M. & Weibelzahl, S. (2009). Log file analysis for disengagement detection in e-Learning environments. *User Modeling and User-Adapted Interaction*, *19*(4), 341–385. doi: 10.1007/s11257-009-9065-5

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd Ed.). Hillsdale: Lawrence Erlbaum Associates.

Cohn, J. & Kanade, T. (2006). Use of automated facial image analysis for measurement of emotion expression. In J. B. Allen & J. A. Coan (Eds.), *The handbook of emotion elicitation and assessment. Oxford University Press Series in Affective Science*. New York: Oxford.

Conati, C., Hoque, E., Toker, D. & Steichen, B. (2013). When to adapt: Detecting user's confusion during visualization processing. In *Proc. of the 1st International Workshop on User-Adaptive Information Visualization (WUAV 2013), in conjunction with the 21st conference on User Modeling, Adaptation and Personalization (UMAP 2013)*.

Cooke, L. (2006). Is the Mouse a " Poor Man ' s Eye Tracker "? In *Society for Technical Communication Conference* (pp. 252–255).

Corrin, L. & de Barba, P. (2014). Exploring students ' interpretation of feedback delivered through learning analytics dashboards. In B. Hegarty, J. McDonald & S.-K. Loke (Eds.), *Rhetoric and Reality: Critical perspectives on educational technology. Proceedings ascilite Dunedin 2014* (pp. 629–633).

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–185. doi: 10.1017/S0140525X01003922

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, *18*(1), 32–80. doi: 10.1109/79.911197

Craig, S. D., D'Mello, S., Witherspoon, A. & Graesser, A. (2008). Emote aloud during learning with AutoTutor: Applying the Facial Action Coding System to cognitive–affective states during learning. *Cognition & Emotion*, *22*(5), 777–788. doi: 10.1080/02699930701516759

Craig, S. D., Gholson, B. & Driscoll, D. M. (2002). Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features and redundancy. *Journal of Educational Psychology*, *94*(2), 428–434. doi: 10.1037//0022-0663.94.2.428

D'Mello, S. & Graesser, A. (2007). Mind and Body: Dialogue and Posture for Affect Detection in Learning Environments. *International Conference on Artificial Intelligence in Education*, 161–168.

D'Mello, S. & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*, 145–157. doi: 10.1016/j.learninstruc.2011.10.001

D'Mello, S., Jackson, T., Craig, S., Morgan, B., Chip, P., White, H., … Graesser, A. (2008). AutoTutor Detects and Responds to Learners Affective and Cognitive States. *Proceedings of the Workshop on Emotional and Cognitive Issues in ITS in Conjunction with the 9th International Conference on Intelligent Tutoring Systems*, *11*, 31–43. doi: 10.1109/TE.2005.856149

D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B. & Graesser, A. (2007). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, *18*(1), 45–80. doi: 10.1007/s11257-007-9037-6

D'Mello, S. K. & Graesser, A. C. (2014). Confusion. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International Handbook of Emotions in Education* (pp. 289–310). London: Routledge.

D'Mello, S. K. & Kory, J. (2015). A Review and Meta-Analysis of Multimodal Affect

Detection Systems. *ACM Computing Surveys*, *47*(3), 1–36. doi: 10.1145/2682899

D'Mello, S., Lehman, B., Pekrun, R. & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, *29*, 153–170. doi: 10.1016/j.learninstruc.2012.05.003

Danner, D., Bluemke, M., Treiber, L., Berres, S., Soto, C. & John, O. (2016). Die deutsche Version des Big Five Inventory (BFI-2). *Zusammenstellung Sozialwissenschaftlicher Items Und Skalen*. doi: 10.6102/zis247

Davis, E. A. (2003). Prompting Middle School Science Students for Productive Reflection: Generic and Directed Prompts. *Journal of the Learning Sciences*, *12*(1), 91–142. doi: 10.1207/S15327809JLS1201_4

Dawson, S., Jovanovic, J., Gašević, D. & Pardo, A. (2017). From prediction to impact. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17*, 474–478. doi: 10.1145/3027385.3027405

de Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P. & Paas, F. (2010). Attention guidance in learning from a complex animation: Seeing is understanding? *Learning and Instruction*, *20*(2), 111–122. doi: 10.1016/j.learninstruc.2009.02.010

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. doi: 10.3758/s13428-014-0458-y

De Lucia, P. R., Preddy, D., Derby, P., Tharanathan, A. & Putrevu, S. (2014). Eye movement behavior during confusion: Toward a method. In *Proceedings of the Human Factors and Ergonomics Society 58th Annual Meeting* (pp. 1300–1304). doi: 10.1177/1541931214581271

Dehn, D. M. & Van Mulken, S. (2000). The impact of animated interface agents: a review of empirical research. *International Journal of Human-Computer Studies*, *52*(1), 1–22. doi: 10.1006/ijhc.1999.0325

Dent, A. L. & Koenka, A. C. (2016). The Relation Between Self-Regulated Learning and Academic Achievement Across Childhood and Adolescence: A Meta-Analysis. *Educational Psychology Review*, *28*(3), 425–474. doi: 10.1007/s10648-015-9320-8

Dignath, C., Buettner, G. & Langfeldt, H. P. (2008). How can primary school students learn self-regulated learning strategies most effectively?. A meta-analysis on self-regulation training programmes. *Educational Research Review*, *3*(2), 101–129. doi:

10.1016/j.edurev.2008.02.003

Dinsmore, D. L., Alexander, P. A. & Loughlin, S. M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review*, *20*(4), 391–409. doi: 10.1007/s10648-008-9083-6

Dirkin, H. K., Mishra, P. & Altermatt, E. (2005). All or nothing: levels of sociability of a pedagogical software agent and its impact on student perceptions and learning. *Journal of Educational Multimedia and Hypermedia*, *14*(2), 113–127.

Domagk, S. (2008). *Pädagogische Agenten in multimedialen Lernumgebungen: empirische Studien zum Einfluss der Sympathie auf Motivation und Lernerfolg*. Berlin: Logos-Verlag.

Domagk, S. (2010). Do pedagogical agents facilitate learner motivation and learning outcomes? *Journal of Media Psychology: Theories, Methods, and Applications*, *22*(2), 84–97. doi: 10.1027/1864-1105/a000011

Drachsler, H. & Greller, W. (2016). Privacy and analytics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16* (pp. 89–98). New York, USA: ACM Press. doi: 10.1145/2883851.2883893

Dunlosky, J. & Lipko, A. R. (2007). Metacomprehension: A Brief History and How to Improve Its Accuraccy. *Current Directions in Psychological Science*, *16*(4), 228–232.

Durso, F. T., Geldbach, K. M. & Corballis, P. (2012). Detecting confusion using facial electromyography. *Human Factors*, *54*(1), 60–69. doi: 10.1177/0018720811428450

Efklides, A. (2008). Metacognition. *European Psychologist*, *13*(4), 277–287. doi: 10.1027/1016-9040.13.4.277

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist*, *46*(1), 6–25. doi: 10.1080/00461520.2011.538645

Ekman, P. & Rosenberg, E. L. (2005). *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. (P. Ekman & E. L. Rosenberg, Eds.). New York, USA: Oxford University Press. doi: 10.1093/acprof:oso/9780195179644.001.0001

Ekstrom, R. B. R., French, J. J. W., Harman, H. H. & Dermen, D. (1976). Manual for kit of factor-referenced cognitive tests. *Princeton NJ Educational Testing Service*, *102*(41),

117. doi: 10.1073/pnas.0506897102

Epp, C., Lippold, M. & Mandryk, R. L. (2011). Identifying emotional states using keystroke dynamics. *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11*, 715–724. doi: 10.1145/1978942.1979046

Erpenbeck, J. & Sauter, W. (2013). *So werden wir lernen: Kompetenzentwicklung in einer Welt fühlender Computer, kluger Wolken und sinnsuchender Netze*. Berlin: Springer.

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 117–140. doi: 10.1177/001872675400700202

Field, A. (2009). *Discovering statistics using IBM SPSS Statistics* (4th editio.). Sage Publications.

Flavell, J. (1979). Metacognition and cognitive monitoring: a new area of cognitive-developemental inquiry. *American Psychologist*, *34*(10), 906–911. doi: 10.1037/0003-066X.34.10.906

Freeman, J. B. & Ambady, N. (2009). Motions of the hand expose the partial and parallel activation of stereotypes: Research report. *Psychological Science*, *20*(10), 1183–1188. doi: 10.1111/j.1467-9280.2009.02422.x

Freeman, J. B. & Ambady, N. (2010). MouseTracker: software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–41. doi: 10.3758/BRM.42.1.226

Freeman, J. B., Pauker, K., Apfelbaum, E. P. & Ambady, N. (2010). Continuous dynamics in the real-time perception of race. *Journal of Experimental Social Psychology*, *46*(1), 179–185. doi: 10.1016/j.jesp.2009.10.002

Gaševic, D., Dawson, S. & Siemens, G. (2015). Let ' s not forget : Learning Analytics are about Learning Course Signals : Lessons Learned. *TechTrends59*, (1), 71–64. doi: 10.1007/s11528-014-0822-x

Gegenfurtner, A., Lehtinen, E., Jarodzka, H. & Säljö, R. (2017). Effects of eye movement modeling examples on adaptive expertise in medical image diagnosis. *Computers & Education*, *113*, 212–225. doi: 10.1016/j.compedu.2017.06.001

Gerjets, P., Scheiter, K. & Cierniak, G. (2009). The Scientific Value of Cognitive Load Theory: A Research Agenda Based on the Structuralist View of Theories. *Educational Psychology Review*, *21*(1), 43–54. doi: 10.1007/s10648-008-9096-1

Glazer, N. (2011). Challenges with graph interpretation: A review of the literature. *Studies in Science Education*, *47*(2), 183–210. doi: 10.1080/03057267.2011.605307

Glogger, I., Holzäpfel, L., Schwonke, R., Nückles, M. & Renkl, A. (2009). Activation of Learning Strategies in Writing Learning Journals. *Zeitschrift Für Pädagogische Psychologie*, *23*(2), 95–104. doi: 10.1024/1010-0652.23.2.95

Glucksberg, S. & McCloskey, M. (1981). Decisions about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning & Memory*, *7*(5), 311–325. doi: 10.1037/0278-7393.7.5.311

Godwin-Jones, R. (2018). Second language writing online: An update. *Language Learning & Technology*, *22*(1), 1–15. doi: 10.1177/001088049403500408

Gosling, S. D., Vazire, S., Srivastava, S. & John, O. P. (2004). Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist*, *59*(2), 93–104. doi: 10.1037/0003-066X.59.2.93

Graesser, A. C. (2011). Learning, thinking, and emoting with discourse technologies. *American Psychologist*, *66*(8), 746–757. doi: 10.1037/a0024974

Graesser, A. C., Lu, S., Olde, B. A., Cooper-Pye, E. & Whitten, S. (2005). Question asking and eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices when the devices break down. *Memory and Cognition*, *33*(7), 1235–1247. doi: 10.3758/BF03193225

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P. & Kreuz, R. (1999). AutoTutor: A simulation of a human tutor. *Cognitive Systems Research*, *1*(1), 35–51. doi: 10.1016/S1389-0417(99)00005-4

Graf, S. & Liu, T.-C. (2010). Analysis of learners' navigational behaviour and their learning styles in an online course. *Journal of Computer Assisted Learning*, *26*(2), 116–131. doi: 10.1111/j.1365-2729.2009.00336.x

Greene, J. A. & Azevedo, R. (2007). A Theoretical Review of Winne and Hadwin's Model of Self-Regulated Learning: New Perspectives and Directions. *Review of Educational Research*, *77*(3), 334–372. doi: 10.3102/003465430303953

Greene, J. A., Robertson, J. & Costa, L.-J. C. (2011). Assessing self-regulated learning using think-aloud methods. *Handbook of Self-Regulation of Learning and Performance*, 313–328.

Grimes, M. & Valacich, J. (2015). Mind Over Mouse: The Effect of Cognitive Load on Mouse Movement Behavior. *ICIS 2015 Proceedings*.

Guo, Q. & Agichtein, E. (2010). Towards predicting web searcher gaze position from mouse movements. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10* (p. 3601). New York, New York, USA: ACM Press. doi: 10.1145/1753846.1754025

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*(2), 149–161. doi: 10.1007/BF02289162

Hadwin, A. F., Järvelä, S. & Miller, M. (2017). Self-Regulated, Co-Regulated, and Socially Shared Regulation of Learning. *Handbook of Self-Regulation of Learning and Performance*, (13109). doi: 10.4324/9780203839010.ch5

Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J. & Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition and Learning*, *2*(2–3), 107–124. doi: 10.1007/s11409-007-9016-7

Harper, S. (2009). Eye Tracking Scanpath Analysis Techniques on Web Pages : A Survey , Evaluation and Comparison. *Journal of Eye Movement Research*, *9*(1), 1–19. doi: 10.16910/jemr.9.1.2

Harting, K. & Erthal, M. (2005). History of Distance Learning. *Information Technology, Learning and Performance Journal*, *23*(1), 35–44. doi: ProQuest ID: 219815808

Hartman, H. (2001). Developing Students' Metacognitive Knowledge and Strategies. In H. Hartman (Ed.), *Metacognition in Learning and Instruction* (pp. 33–68). New York: City University of New York. doi: 10.1007/978-94-017-2243-8

Hauger, D., Paramythis, A. & Weibelzahl, S. (2011). Using browser interaction data to determine page reading behavior. In J. A. Konstan, R. Conejo, J. L. Marzo & N. Oliver (Eds.), *User Modelling, Adaption and Personalization - UMAP 2011* (pp. 147–158).

Heidig, S. & Clarebout, G. (2011). Do pedagogical agents make a difference to student motivation and learning? *Educational Research Review*, *6*(1), 27–54. doi: 10.1016/j.edurev.2010.07.004

Henrich, J., Heine, S. J. & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. doi: 10.1017/S0140525X0999152X

Henriques, R., Paiva, A. & Antunes, C. (2013). Accessing emotion patterns from affective interactions using electrodermal activity. *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 43–48. doi: 10.1109/ACII.2013.14

Hester, R., Barre, N., Murphy, K., Silk, T. J. & Mattingley, J. B. (2008). Human medial frontal cortex activity predicts learning from errors. *Cerebral Cortex*, *18*(8), 1933–1940. doi: 10.1093/cercor/bhm219

Hilbig, B. E. (2015). Reaction time effects in lab- versus Web-based research: Experimental evidence. *Behavior Research Methods*, 1718–1724. doi: 10.3758/s13428-015-0678-9

Horwitz, R. (2013). *Classifying mouse movements and providing help in web surveys*. Retrieved from http://drum.lib.umd.edu/bitstream/handle/1903/14039/Horwitz_umd_0117E_14107.pdf?sequence=1&isAllowed=y

Horwitz, R., Kieslich, P. J. & Kreuter, F. (2017). Learning from Mouse Movements: Improving Questionnaire and Respondents' User Experience through Passive Data Collection. *IAB-Discussion Paper*, *34*.

Horwitz, R., Kreuter, F. & Conrad, F. (2017). Using Mouse Movements to Predict Web Survey Response Difficulty. *Social Science Computer Review*, *35*(3), 388–405. doi: 10.1177/0894439315626360

Huang, J. & White, R. (2012). User See, User Point: Gaze and Cursor Alignment in Web Search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1341–1350. doi: 10.1145/2207676.2208591

Huang, J., White, R. W. & Dumais, S. (2011). No clicks, no problem: Using cursor movements to understand and improve search. *Proceedings of the 29th SIGCHI Conference on Human Factors in Computing Systems*, 1225. doi: 10.1145/1978942.1979125

Hudlicka, E. & McNeese, M. D. (2002). Assessment of User Affective and Belief States for Interface Adaptation: Application to an Air Force Pilot Task. *User Modeling and User-Adapted Interaction*, *12*(1), 1–47. doi: 10.1023/A:1013337427135

Hussain, M. S., AlZoubi, O., Calvo, R. A. & D'Mello, S. K. (2011). Affect Detection from Multichannel Physiology during Learning Sessions with AutoTutor. In G. Biswas, S. Bull, J. Kay & A. Mitrovic (Eds.), *Artificial Intelligence in Education* (pp. 131–138).

Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-21869-9_19

Jarodzka, H., van Gog, T., Dorr, M., Scheiter, K. & Gerjets, P. (2013). Learning to see: Guiding students' attention via a Model's eye movements fosters learning. *Learning and Instruction*, *25*, 62–70. doi: 10.1016/j.learninstruc.2012.11.004

Järvelä, S., Volet, S. & Järvenoja, H. (2010). Research on motivation in collaborative learning: Moving beyond the cognitive-situative divide and combining individual and social processes. *Educational Psychologist*, *45*(1), 15–27. doi: 10.1080/00461520903433539

Jarvenpaa, S. L. (1989). The Effect of Task Demands and Graphical Format on Information Processing Strategies. *Management Science*, *35*(3), 285–303. doi: 10.1287/mnsc.35.3.285

Jivet, I., Scheffel, M., Drachsler, H. & Specht, M. (2017). Awareness Is Not Enough: Pitfalls of Learning Analytics Dashboards in the Educational Practice. In É. Lavoué, H. Drachsler, K. Verbert, J. Broisin & M. Pérez-Sanagustín (Eds.), *12th European Conference on Technology Enhanced Learning, EC-TEL 2017* (Vol. 10474, pp. 82–96). Basel: Springer International Publishing AG. doi: 10.1007/978-3-319-66610-5_7

Johnson, W., Shaw, E. & Ganeshan, R. (1998). Pedagogical agents on the web. *ITS Workshop on Pedagogical Agents*.

Jones, S. & Burnett, G. (2008). Spatial ability and learning to program. *Human Technology*, *4*(May), 47–61. doi: 10.17011/ht/urn.200804151352

Jorgensen, Z. & Yu, T. (2011). On mouse dynamics as a behavioral biometric for authentication. In *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security - ASIACCS '11* (p. 476). New York, New York, USA: ACM Press. doi: 10.1145/1966913.1966983

Kalyuga, S. (2008). When less is more in cognitive diagnosis: A rapid online method for diagnosing learner task-specific expertise. *Journal of Educational Psychology*, *100*(3), 603–612. doi: 10.1037/0022-0663.100.3.603

Kalyuga, S. (2011). Cognitive Load Theory: How Many Types of Load Does It Really Need? *Educational Psychology Review*, *23*(1), 1–19. doi: 10.1007/s10648-010-9150-7

Keim, D., Andrienko, G., Fekete, J., Görg, C., Melançon, G., Keim, D., … Stasko, J. T. (2008). Visual Analytics: Definition, Process and Challenges. In A. Kerren, J. T. Stasko,

J. Fekete & C. North (Eds.), *Information Visualization - Human-Centered Issues and Perspectives* (pp. 154–175). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Kieslich, P. J. & Henninger, F. (2017). Mousetrap: An integrated, open-source mouse-tracking package. *Behavior Research Methods*, *49*(5), 1652–1667. doi: 10.3758/s13428-017-0900-z

Király, S., Nehéz, K. & Hornyák, O. (2017). Some aspects of grading Java code submissions in MOOCs. *Research in Learning Technology*, *25*(1063519), 1–16. doi: 10.25304/rlt.v25.1945

Kirschner, P. a. (2002). Cognitive load theory: Implications of cognitive load theory on the design of learning. *Learning and Instruction*, *12*(1), 1–10. doi: 10.1016/S0959-4752(01)00014-7

Kirschner, P. A., Sweller, J. & Clark, R. E. (2006). Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educational Psychologist*, *41*(2), 75–86. doi: 10.1207/s15326985ep4102_1

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12. doi: 10.1177/2053951714528481

Klepsch, M., Schmitz, F. & Seufert, T. (2017). Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load. *Frontiers in Psychology*, *8*, 1997. doi: 10.3389/fpsyg.2017.01997

Kolakowska, A. (2013). A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *2013 6th International Conference on Human System Interactions (HSI)* (pp. 548–555). IEEE. doi: 10.1109/HSI.2013.6577879

Koriat, A., Ma'ayan, H. & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, *135*(1), 36–69. doi: 10.1037/0096-3445.135.1.36

Krebs, M.-C., Schüler, A. & Scheiter, K. (2018). Just follow my eyes: The influence of model-observer similarity on Eye Movement Modeling Examples. *Learning and Instruction*. doi: 10.1016/j.learninstruc.2018.10.005

Krohne, H. W., Egloff, B., Kohlmann, C.-W. & Tausch, A. (1996). Untersuchungen mit einer deutschen Form der Positive and Negative Affect Schedule (PANAS). *Diagnostica*, (42), 139–156.

Kucukyilmaz, T. & Cambazoglu, B. (2006). Advances in Information Systems. In T. Yakhno & E. J. Neuhold (Eds.), *ADVIS 2006* (Vol. 4243, pp. 274–283). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/11890393

Lachner, A., Burkhart, C. & Nückles, M. (2017). Formative computer-based feedback in the university classroom: Specific concept maps scaffold students' writing. *Computers in Human Behavior*, *72*, 459–469. doi: 10.1016/j.chb.2017.03.008

Lahl, O. & Pietrowsky, R. (2008). Tracer: a general-purpose software library for logging events in computerized experiments. *Behavior Research Methods*, *40*(4), 1163–9. doi: 10.3758/BRM.40.4.1163

Lali, P., Naghizadeh, M., Nasrollahi, H., Moradi, H. & Mirian, M. S. (2014). Your mouse can tell about your emotions. In *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)* (pp. 47–51). IEEE. doi: 10.1109/ICCKE.2014.6993360

Lallé, S., Conati, C. & Carenini, G. (2016). Predicting confusion in information visualization from eye tracking and interaction data. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2529–2535.

Landis, J. R. & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. doi: 10.2307/2529310

Larzelere, R. E., Kuhn, B. R. & Johnson, B. (2004). The Intervention Selection Bias: An Underrecognized Confound in Intervention Research. *Psychological Bulletin*, *130*(2), 289–303. doi: 10.1037/0033-2909.130.2.289

Leijten, M. & Van Waes, L. (2013). Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication*, *30*(3), 358–392. doi: 10.1177/0741088313491692

Leijten, M., Van Waes, L. & Ransdell, S. (2010). Correcting Text Production Errors: Isolating the Effects of Writing Mode From Error Span, Input Mode, and Lexicality. *Written Communication*, *27*(2), 189–227. doi: 10.1177/0741088309359139

Leiva, L. A. & Hernando, R. V. (2007). (Smt) Real Time Mouse Tracking Registration and Visualization Tool for Usability Evaluation on Websites. *Proceedings of the IADIS*

International Conference on WWW/Internet, 187–192.

Leiva, L. A. & Vivó, R. (2013). Web browsing behavior analysis and interactive hypervideo. *ACM Transactions on the Web*, *7*(4), 1–28. doi: 10.1145/2529995.2529996

Leong, F. H. (2016). Fine-Grained Detection of Programming Students' Frustration Using Keystrokes, Mouse Clicks and Interaction Logs. *Open Journal of Social Sciences*, *4*, 9–18. doi: 10.4236/jss.2016.49002

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A. & Bhogal, R. S. (1997). The persona effect: affective impact of animated pedagogical agents. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems* (pp. 359–366). ACM.

Lim, Y. M., Ayesh, A. & Stacey, M. (2014). Detecting cognitive stress from keyboard and mouse dynamics during mental arithmetic. *Proceedings of 2014 Science and Information Conference, SAI 2014*, 146–152. doi: 10.1109/SAI.2014.6918183

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J. & Bartlett, M. (2011). The computer expression recognition toolbox (CERT). In *Face and Gesture 2011* (pp. 298–305). IEEE. doi: 10.1109/FG.2011.5771414

Lloyd, S. A. & Robertson, C. L. (2012). Screencast Tutorials Enhance Student Learning of Statistics. *Teaching of Psychology*, *39*(1), 67–71. doi: 10.1177/0098628311430640

Lusk, M. M. & Atkinson, R. K. (2007). Animated pedagogical agents: Does their degree of embodiment impact learning from static or animated worked examples? *Applied Cognitive Psychology*, *21*(6), 747–764.

Maehr, W. (2005). *eMotion: Estimation of User's Emotional State by Mouse Motions*. Fachhochschule Vorarlberg.

Malmberg, J., Järvelä, S., Holappa, J., Haataja, E., Huang, X. & Siipo, A. (2018). Going beyond what is visible: What multichannel data can reveal about interaction in the context of collaborative learning? *Computers in Human Behavior*, *2018*. doi: 10.1016/j.chb.2018.06.030

Mandler, G. (1990). Interruption (discrepancy) theory: Review and extensions. In S. Fisher & C. L. Cooper (Eds.), *On the move: The psychology of change and Transition* (pp. 13–32). Chichester: Wiley.

Martín-Gutiérrez, J., Mora, C. E., Añorbe-Díaz, B. & González-Marrero, A. (2017). Virtual technologies trends in education. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(2), 469–486. doi: 10.12973/eurasia.2017.00626a

Mason, L., Pluchino, P. & Tornatora, M. C. (2015). Eye-movement modeling of integrative reading of an illustrated text: Effects on processing and learning. *Contemporary Educational Psychology*, *41*, 172–187. doi: 10.1016/j.cedpsych.2015.01.004

Mayer, R. E. (1975). Information Processing Variables in Learning to Solve Problems. *Review of Educational Research*, *45*(4), 525–541. doi: 10.3102/00346543045004525

Mayer, R. E. (1981). The Psychology of How Novices Learn Computer Programming. *ACM Computing Surveys*, *13*(1), 121–141. doi: 10.1145/356835.356841

Mayer, R. E. (2005). Principles of Multimedia Learning Based on Social Cues : Personalization, Voice, and Image Principles. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 201–212). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511816819.014

Mayer, R. E. (2009). *Multimedia Learning* (2nd ed.). New York, NY, USA: Cambridge University Press.

Mayer, R. E. & DaPra, C. S. (2012). An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology: Applied*, *18*(3), 239–52. doi: 10.1037/a0028616

Mayer, R. E., Dow, G. T. & Mayer, S. (2003). Multimedia learning in an interactive self-explaining environment: What works in the design of agent-based microworlds? *Journal of Educational Psychology*, *95*(4), 806.

Mayer, R. E. & Moreno, R. (2003). Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educational Psychologist*, *38*(1), 43–52. doi: 10.1207/S15326985EP3801_6

McQuiggan, S. W., Mott, B. W. & Lester, J. C. (2008). Modeling self-efficacy in intelligent tutoring systems: An inductive approach. *User Modeling and User-Adapted Interaction*, *18*(1–2), 81–123. doi: 10.1007/s11257-007-9040-y

Miller, A. (2004). Video-Cued Recall: Its use in a Work Domain Analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(15), 1643–1647. doi: 10.1177/154193120404801503

Miller, B. W. (2015). Using Reading Times and Eye-Movements to Measure Cognitive Engagement. *Educational Psychologist*, *50*(March), 31–42. doi:

10.1080/00461520.2015.1004068

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limites on out Capacity for Processing Information. *Psychological Review*, *65*(2), 81–97. doi: http://dx.doi.org/10.1037/h0043158

MMB-Trendmonitor. (2014). *Wenn der digitale Lernassistent uns an die Hand nimmt*. Essen: MMB-Institut für Medien- und Kompetenzforschung.

Moreno, R. (2003). Animated pedagogical agents: how do they help students construct knowledge from interactive multimedia environments. *Unpublished Manuscript*.

Moreno, R. (2004). Decreasing cognitive load for novice students: effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science*, *32*(1/2), 99–113. doi: 10.1023/B:TRUC.0000021811.66966.1d

Moreno, R. (2005). Multimedia learning with animated pedagogical agents. In R. E. Mayer (Ed.), *Cambridge Handbook of multimedia Learning* (pp. 507–524). New York: Cambridge University Press.

Moreno, R. & Mayer, R. E. (2005). Role of guidance, reflection, and interactivity in an agent-based multimedia game. *Journal of Educational Psychology*, *97*(1), 117–128. doi: 10.1037/0022-0663.97.1.117

Moreno, R., Mayer, R. E., Spires, H. A. & Lester, J. C. (2001). The case for social agency in computer-based teaching: do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction*, *19*(2), 177–213. doi: 10.1207/S1532690XCI1902_02

Moreno, R. & Park, B. (2010). Cognitive load theory: Historical development and relation to other theories. In J. L. Plass, R. Moreno & R. Brünken (Eds.), *Cognitive load theory* (pp. 9–28). Cambridge University Press New York, NY.

Morris, S. B. (2008). Estimating Effect Sizes From Pretest-Posttest-Control Group Designs. *Organizational Research Methods*, *11*(2), 364–386. doi: 10.1177/1094428106291059

Mozilla Development Network. (2015). MutationObserver. Retrieved February 10, 2017, from https://developer.mozilla.org/de/docs/Web/API/MutationObserver

Mozilla Development Network. (2016). AJAX. Retrieved February 10, 2017, from https://developer.mozilla.org/de/docs/AJAX

Mueller, F. & Lockerd, A. (2001). Cheese: tracking mouse movement activity on websites, a tool for user modeling. In *CHI'01 extended abstracts on Human factors* (pp. 279–280).

Münzer, S., Seufert, T. & Brünken, R. (2009). Learning from multimedia presentations: Facilitation function of animations and spatial abilities. *Learning and Individual Differences*, *19*(4), 481–485. doi: 10.1016/j.lindif.2009.05.001

Murphy, P. K. & Alexander, P. A. (2000). A motivated exploration of motivation terminology. *Contemporary Educational Psychology*, *25*(1), 3–53. doi: 10.1006/ceps.1999.1019

Narciss, S. & Huth, K. (2004). How to design informative tutoring feedback. In H. M. Niegemann, D. Leutner & R. Brünken (Eds.), *Instruction Design for Multimedia Learning* (pp. 181–195). New York: Waxmann Verlag.

Nass, C., Moon, Y., Fogg, B. J., Reeves, B. & Dryer, D. C. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, *43*(2), 223–239. doi: 10.1006/ijhc.1995.1042

Nazemi, K., Burkhardt, D., Hoppe, D., Nazemi, M. & Kohlhammer, J. (2015). Web-based Evaluation of Information Visualization. *Procedia Manufacturing*, *3*, 5527–5534. doi: 10.1016/j.promfg.2015.07.718

Nelson, T. O. (1990). Metamemory: A Theoretical Framework and New Findings. *The Psychology of Learning and Motivation*, *26*, 125–173. doi: 10.1016/S0079-7421(08)60053-5

Nijstad, B. A., Stroebe, W. & Lodewijkx, H. F. M. (1999). Persistence of Brainstorming Groups: How Do People Know When to Stop? *Journal of Experimental Social Psychology*, *35*(2), 165–185. doi: 10.1006/jesp.1998.1374

Nückles, M., Hübner, S. & Renkl, A. (2009). Enhancing self-regulated learning by writing learning protocols. *Learning and Instruction*, *19*(3), 259–271. doi: 10.1016/j.learninstruc.2008.05.002

Paas, F. & Ayres, P. (2014). Cognitive Load Theory: A Broader View on the Role of Memory in Learning and Education. *Educational Psychology Review*, *26*(2), 191–195. doi: 10.1007/s10648-014-9263-5

Palinko, O., Kun, A. L., Shyrokov, A. & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on*

Eye-Tracking Research & Applications - ETRA '10 (Vol. 287, p. 141). New York, New York, USA: ACM Press. doi: 10.1145/1743666.1743701

Panadero, E. (2017). A Review of Self-regulated Learning: Six Models and Four Directions for Research. *Frontiers in Psychology*, *8*, 422. doi: 10.3389/fpsyg.2017.00422

Pardo, A., Poquet, O., Martinez-Maldonado, R. & Dawson, S. (2017). Provision of Data-Driven Student Feedback in LA & EDM. In C. Lang, G. Siemens, A. Wise & D. Gašević (Eds.), *Handbook of Learning Analytics* (pp. 163–174). Society for Learning Analytics Research (SoLAR). doi: 10.18608/hla17.014

Pardos, Z. A., Baker, R. S. J. ., Pedro, M. S., Gowda, S. M. & Gowda, S. M. (2014). Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. *Journal of Learning Analytics*, *1*(1), 107–128. doi: 10.1145/2460296.2460320

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. doi: 10.1002/acp.3100

Park, Y. & Jo, I. (2015). Development of the Learning Analytics Dashboard to Support Students ' Learning Performance Learning Analytics Dashboards (LADs). *Journal of Universal Computer Science*, *21*(1), 110–133. doi: 10.3217/jucs-021-01-0110

Pecchinenda, A. & Smith, C. . (1996). The Affective Significance of Skin Conductance Activity During a Difficult Problem-solving Task. *Cognition & Emotion*, *10*(5), 481–504. doi: 10.1080/026999396380123

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*(4), 315–341. doi: 10.1007/s10648-006-9029-9

Pekrun, R. & Stephens, E. J. (2016). Academic emotions. In K. R. Harris, S. Graham & T. Urdan (Eds.), *APA educational psychology handbook, Vol 2: Individual differences and cultural and contextual factors.* (pp. 3–31). Washington: American Psychological Association. doi: 10.1037/13274-001

Pentel, A. (2015). Employing think-aloud protocol to connect user emotions and mouse movements. In *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1–5). IEEE. doi: 10.1109/IISA.2015.7387970

Perry, N. E. & Winne, P. H. (2006). Learning from learning kits: gStudy traces of students' self-regulated engagements with computerized content. *Educational Psychology Review*, *18*(3), 211–228. doi: 10.1007/s10648-006-9014-3

Piaget, J. (1928). *Judgment and reasoning in the child*. New York: Harcourt Brace.

Pieger, E. & Bannert, M. (2018). Differential effects of students' self-directed metacognitive prompts. *Computers in Human Behavior*, *86*, 165–173. doi: 10.1016/j.chb.2018.04.022

Pijeira-Díaz, H. J., Drachsler, H., Järvelä, S. & Kirschner, P. A. (2016). Investigating collaborative learning success with physiological coupling indices based on electrodermal activity. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16* (pp. 64–73). New York, New York, USA: ACM Press. doi: 10.1145/2883851.2883897

Pintrich, P. R. (2000). The Role of Goal Orientation in Self-Regulated Learning. In *Handbook of Self-Regulation* (pp. 451–502). Elsevier. doi: 10.1016/B978-012109890-2/50043-3

Pirrie, A. & Thoutenhoofd, E. D. (2013). Learning to learn in the European Reference Framework for lifelong learning. *Oxford Review of Education*, *39*(5), 609–626. doi: 10.1080/03054985.2013.840280

Plass, J. L., Heidig, S., Hayward, E. O., Homer, B. D. & Um, E. (2014). Emotional design in multimedia learning: Effects of shape and color on affect and learning. *Learning and Instruction*, *29*, 128–140. doi: 10.1016/j.learninstruc.2013.02.006

Plass, J. L., Moreno, R. & Brünken, R. (2010). *Cognitive Load Theory*. New York: Cambridge University Press.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y. & Podsakoff, N. P. (2003). Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *Journal of Applied Psychology*, *88*(5), 879–903. doi: 10.1037/0021-9010.88.5.879

Postma-Nilsenová, M., Postma, E. & Tates, K. (2015). Automatic Detection of Confusion in Elderly Users of a Web-Based Health Instruction Video. *Telemedicine and E-Health*, *21*(6), 514–519. doi: 10.1089/tmj.2014.0061

Puntambekar, S., Sullivan, S. A. & Hübscher, R. (2013). Analyzing Navigation Patterns to Scaffold Metacognition in Hypertext Systems. In R. Azevedo (Ed.), *International Handbook of Metacognition and Learning Technologies* (pp. 261–275). doi:

10.1007/978-1-4419-5546-3_18

Puustinen, M. & Pulkkinen, L. (2001). Models of Self-regulated Learning: A review. *Scandinavian Journal of Educational Research*, *45*(3), 269–286. doi: 10.1080/00313830120074206

Ramos-Soto, A., Lama, M., Vazquez-Barreiros, B., Bugarin, A. & Barro, M. M. S. (2015). Towards Textual Reporting in Learning Analytics Dashboards. In *2015 IEEE 15th International Conference on Advanced Learning Technologies* (pp. 260–264). IEEE. doi: 10.1109/ICALT.2015.96

Rani, P., Sarkar, N. & Smith, C. . (2003). Affect-sensitive human-robot cooperation - theory and experiments. In *2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422)* (Vol. 2, pp. 2382–2387). IEEE. doi: 10.1109/ROBOT.2003.1241949

Razak, A. M. R. & Ali, M. A. Z. (2016). Instructional screencast: A research conceptual framework. *Turkish Online Journal of Distance Education*, *17*(2), 74–87. doi: 10.17718/tojde.21316

Reimann, P., Markauskaite, L. & Bannert, M. (2014). e-Research and learning theory: What do sequence and process mining methods contribute? *British Journal of Educational Technology*, *45*(3), 528–540. doi: 10.1111/bjet.12146

Reips, U.-D. (2000). The Web Experiment Method. In *Psychological Experiments on the Internet* (pp. 89–117). Elsevier. doi: 10.1016/B978-012099980-4/50005-8

Reips, U.-D. & Birnbaum, M. H. (2011). Behavioral Research and Data Collection via the Internet. In K.-P. L. Vu & R. W. Proctor (Eds.), *Handbook of human factors in Web design* (2nd ed., pp. 563–585). Mahwah, New Jersey: Erlbaum.

Reips, U.-D. & Neuhaus, C. (2002). WEXTOR: a Web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods, Instruments, & Computers : A Journal of the Psychonomic Society, Inc*, *34*(2), 234–240. doi: 10.3758/BF03195449

Reips, U.-D. & Stieger, S. (2004). Scientific LogAnalyzer: A Web-based tool for analyses of server log files in psychological research. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 304–311. doi: 10.3758/BF03195576

Reips, U. D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, *49*(4), 243–256. doi: 10.1026//1618-3169.49.4.243

Renkl, A., Skuballa, I. T., Schwonke, R., Harr, N. & Leber, J. (2015). The Effects of Rapid Assessments and Adaptive Restudy Prompts in Multimedia Learning. *Educational Technology & Society*, *18*(4), 184–198.

Rheem, H., Verma, V. & Becker, D. V. (2018). Use of Mouse-tracking Method to Measure Cognitive Load. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 1982–1986. doi: 10.1177/1541931218621449

Rheinberg, F., Vollmeyer, R. & Burns, B. D. (2001). QCM : A questionnaire to assess current motivation in learning situations. *Diagnostica*, *47*(2), 57–66. doi: 10.1026//0012-1924.47.2.57

Rheinberg, F., Vollmeyer, R. & Rollett, W. (2000). Motivation and Action in Self-Regulated Learning. In M. Boekaerts, P. R. Pintrich & M. Zeidner (Eds.), *Handbook of Self-Regulation* (Vol. 1, pp. 503–529). Cambridge: Elsevier. doi: 10.1016/B978-012109890-2/50044-5

Rodrigues, M., Gonçalves, S., Carneiro, D., Novais, P. & Fdez-Riverola, F. (2013). Keystrokes and Clicks: Measuring Stress on E-learning Students. In J. Casillas, F. J. Martínez-López, R. Vicari & F. De la Prieta (Eds.), *Management Intelligent Systems* (Vol. 220, pp. 119–126). Heidelberg: Springer International Publishing. doi: 10.1007/978-3-319-00569-0_15

Salmeron-Majadas, S., Santos, O. C. & Boticario, J. G. (2014). An Evaluation of Mouse and Keyboard Interaction Indicators towards Non-intrusive and Low Cost Affective Modeling in an Educational Context. *Procedia Computer Science*, *35*, 691–700. doi: 10.1016/j.procs.2014.08.151

Scheiter, K., Schubert, C. & Schüler, A. (2017). Self-regulated learning from illustrated text: Eye movement modelling to support use and regulation of cognitive processes during learning from multimedia. *British Journal of Educational Psychology*, *88*(1), 80–94. doi: 10.1111/bjep.12175

Schiefele, U. & Wild, K. P. (1994). Lernstrategien im Studium: Ergebnisse zur Faktorenstruktur und Reliabilität eines neuen Fragebogens [Learning strategies of university students: Factor structure and reliability of a new questionnaire]. *Zeitschrift Für Differentielle Und Diagnostische Psychologie*, *15*, 185–200.

Schipolowski, S., Wilhelm, O., Schroeders, U., Kovaleva, A., Kemper, C. J. & Rammstedt, B. (2013). BEFKI GC-K: Eine Kurzskala zur Messung kristalliner Intelligenz. *Methoden, Daten, Analysen*, *7*(2), 153–181. doi: 10.12758/mda.2013.010

Schipolowski, S., Wilhelm, O., Schroeders, U., Kovaleva, A., Kemper, C. J. & Rammstedt, B. (2014). Kurzskala kristalline Intelligenz (BEFKI GC-K). *Zusammenstellung Sozialwissenschaftlicher Items Und Skalen*. doi: 10.6102/zis220

Schmeck, A., Opfermann, M., van Gog, T., Paas, F. & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: differences between immediate and delayed ratings. *Instructional Science*, *43*(1), 93–114. doi: 10.1007/s11251-014-9328-3

Schnotz, W. & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction*, *13*(2), 141–156. doi: 10.1016/S0959-4752(02)00017-8

Schnotz, W. & Kürschner, C. (2007). A Reconsideration of Cognitive Load Theory. *Educational Psychology Review*, *19*(4), 469–508. doi: 10.1007/s10648-007-9053-4

Schnotz, W., Seufert, T. & Bannert, M. (2001). Lernen mit Multimedia: Pädagogische Verheißungen aus kognitionspsychologischer Sicht. *Psychologie 2000*, 457–467.

Schönbrodt, F. D. & Asendorpf, J. B. (2011). The challenge of constructing psychologically believable agents. *Journal of Media Psychology: Theories, Methods, and Applications*, *23*(2), 100–107. doi: 10.1027/1864-1105/a000040

Schoor, C., Bannert, M. & Brünken, R. (2012). Role of dual task design when measuring cognitive load during multimedia learning. *Educational Technology Research and Development*, *60*(5), 753–768. doi: 10.1007/s11423-012-9251-8

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science*, *26*(1), 113–125. doi: 10.1023/A:1003044231033

Schwarz, N. (2010). Meaning in Context: Metacogtnitive Experiences. In B. Mesquita, L. F. Barrett & E. R. Smith (Eds.), *The Mind in Context* (pp. 105–125). New York: Guilford.

Schwendimann, B. A., Rodriguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., … Dillenbourg, P. (2017). Perceiving Learning at a Glance: A Systematic Literature Review of Learning Dashboard Research. *IEEE Transactions on Learning Technologies*, *10*(1), 30–41. doi: 10.1109/TLT.2016.2599522

Seaman, J. E., Allen, I. E. & Seaman, J. (2018). *Grade Increase: Tracking Distance Education in the United States*. New York: Pearson.

Semmelmann, K. & Weigelt, S. (2017). Online psychophysics: reaction time effects in cognitive experiments. *Behavior Research Methods*, *49*(4), 1241–1260. doi: 10.3758/s13428-016-0783-4

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Troster, G. & Ehlert, U. (2010). Discriminating Stress From Cognitive Load Using a Wearable EDA Device. *IEEE Transactions on Information Technology in Biomedicine*, *14*(2), 410–417. doi: 10.1109/TITB.2009.2036164

Seufert, T. (2003). Supporting coherence formation in learning from multiple representations. *Learning and Instruction*, *13*(2), 227–237. doi: 10.1016/S0959-4752(02)00022-1

Seufert, T. (2018). The interplay between self-regulation in learning and cognitive load. *Educational Research Review*, *24*, 116–129. doi: 10.1016/j.edurev.2018.03.004

Shah, D. (2015). MOOCs in 2015: Breaking Down the Numbers | EdSurge News. Retrieved August 4, 2018, from https://www.edsurge.com/news/2015-12-28-moocs-in-2015-breaking-down-the-numbers

Shea, R., Liu, J., Ngai, E. & Cui, Y. (2013). Cloud gaming: Architecture and performance. *IEEE Network*, *27*(4), 16–21. doi: 10.1109/MNET.2013.6574660

Shen, L., Wang, M. & Shen, R. (2009). Affective eLearning Using “Emotional” Data to Improve Learning in Pervasive Learning Environment. *Educational Technology & Society*, *12*(2), 176–189.

Shiban, Y., Schelhorn, I., Jobst, V., Hörnlein, A., Puppe, F., Pauli, P. & Mühlberger, A. (2015). The appearance effect: Influences of virtual agent features on performance and motivation. *Computers in Human Behavior*, *49*, 5–11. doi: 10.1016/j.chb.2015.01.077

Shute, V. & Zapata-Rivera, D. (2008). Adaptive educational systems. *Adaptive Technologies for Training and Education*, (1), 7–27. doi: 10.1017/CBO9781139049580.004

Silvia, P. J. (2010). Confusion and interest: The role of knowledge emotions in aesthetic experience. *Psychology of Aesthetics, Creativity, and the Arts*, *4*(2), 75–80. doi: 10.1037/a0017081

Simkin, D. & Hastie, R. (1987). An Information-Processing Analysis of Graph Perception. *Journal of the American Statistical Association*, *82*(398), 454–465. doi:

10.1080/01621459.1987.10478448

Singer, L. M. & Alexander, P. A. (2017). Reading Across Mediums: Effects of Reading Digital and Print Texts on Comprehension and Calibration. *Journal of Experimental Education*, *85*(1), 155–172. doi: 10.1080/00220973.2016.1143794

Skitka, L. J. & Sargis, E. G. (2006). The Internet as Psychological Laboratory. *Annual Review of Psychology*, *57*(1), 529–555. doi: 10.1146/annurev.psych.57.102904.190048

Snow, R. E. (1989). Aptitude-treatment interaction as a framework for research on learning and individual differences. *Learning and Individual Differences*, *59*(2), 13–59.

Sonnenberg, C. & Bannert, M. (2016). Evaluating the Impact of Instructional Support Using Data Mining and Process Mining: A Micro-Level Analysis of the Effectiveness of Metacognitive Prompts. *Journal of Educational Data Mining*, *8*(2), 51–83.

Sonnenberg, C. & Bannert, M. (2018). Using Process Mining to examine the sustainability of instructional support: How stable are the effects of metacognitive prompting on self-regulatory behavior? *Computers in Human Behavior*, *2018*. doi: 10.1016/j.chb.2018.06.003

Sottilare, R. A. & Proctor, M. (2012). Passively Classifying Student Mood and Performance within Intelligent Tutors. *Educational Technology & Society*, *15*(2), 101–114.

Špakov, O. & Miniotas, D. (2007). Visualization of Eye Gaze Data using Heat Maps. *Elektronika Ir Elektrotechnika*, *74*(2), 55–58.

Spelman Miller, K. (2000). Academic writers on-line: investigating pausing in the production of text. *Language Teaching Research*, *4*(2), 123–148. doi: 10.1191/136216800675510135

Stifterverband für die Deutsche Wissenschaft e.V. (2017). Höhere Chancen durch höhere Bildung?

Sullins, J. & Graesser, A. C. (2014). The relationship between cognitive disequilibrium, emotions and individual differences on student question generation. *International Journal of Learning Technology*, *9*(3), 221. doi: 10.1504/IJLT.2014.065749

Sullivan, G. M. & Artino, A. R. (2013). Analyzing and Interpreting Data From Likert-Type Scales. *Journal of Graduate Medical Education*, *5*(4), 541–542. doi: 10.4300/JGME-5-4-18

Suneetha, K. R. & Krishnamoorthi, R. (2009). Identifying User Behavior by Analyzing Web Server Access Log File. *International Journal of Computer Science and Network Security*, *9*(4), 327–332. doi: 10.5121/ijnsa.2011.3107

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. doi: 10.1016/0364-0213(88)90023-7

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312. doi: 10.1016/0959-4752(94)90003-5

Sweller, J. (2004). Instructional Design Consequences of an Analogy between Evolution by Natural Selection and Human Cognitive Architecture. *Instructional Science*, *32*, 9–31.

Sweller, J. (2005). Implications of Cognitive Load Theory for Multimedia Learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 19–48). New York: Cambridge University Press.

Sweller, J. (2009). Cognitive bases of human creativity. *Educational Psychology Review*, *21*(1), 11–19. doi: 10.1007/s10648-008-9091-6

Sweller, J., Ayres, P. & Kalyuga, S. (2011). *Cognitive Load Theory*. New York, NY: Springer New York. doi: 10.1007/978-1-4419-8126-4

Sweller, J. & Chandler, P. (1994). Why Some Material Is Difficult to Learn. *Cognition and Instruction*, *12*(3), 185–233. doi: 10.1207/s1532690xci1203_1

Sweller, J., Van Merrienboer, J. J. G. & Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review*, *10*(3), 251–296. doi: 10.1023/A:1022193728205

Tabachnick, B. G. & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.

Tang, J. C., Liu, S. B., Muller, M., Lin, J. & Drews, C. (2006). Unobtrusive but invasive. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work - CSCW '06* (p. 479). New York, New York, USA: ACM Press. doi: 10.1145/1180875.1180948

Teasley, S. D. (2017). Student Facing Dashboards: One Size Fits All? *Technology, Knowledge and Learning*, *22*(3), 377–384. doi: 10.1007/s10758-017-9314-3

Thillmann, H., Künsting, J., Wirth, J. & Leutner, D. (2009). Is it Merely a Question of

"What" to Prompt or Also "When" to Prompt? *Zeitschrift Für Pädagogische Psychologie*, *23*(2), 105–115. doi: 10.1024/1010-0652.23.2.105

Touré-Tillery, M. & Fishbach, A. (2014). How to Measure Motivation: A Guide for the Experimental Social Psychologist. *Social and Personality Psychology Compass*, *8*(7), 328–341. doi: 10.1111/spc3.12110

Trewin, S. (1998). Input logger: General-purpose logging of keyboard and mouse events on an Apple Macintosh. *Behavior Research Methods, Instruments, & Computers*, *30*(2), 327–331.

Tsoulouhas, G., Georgiou, D. & Karakos, A. (2011). Detection of Learner's Affetive State Based on Mouse Movements. *Journal of Computing*, *3*(11), 9–18.

Ullrich, C. & Melis, E. (2002). The Poor Man' s Eyetracker Tool of ActiveMath. In *Proceedings of the World Conference on E-Learning in Corporate Government Healthcare and Higher Education* (pp. 2313–2316).

Um, E. "Rachel," Plass, J. L., Hayward, E. O. & Homer, B. D. (2012). Emotional design in multimedia learning. *Journal of Educational Psychology*, *104*(2), 485–498. doi: 10.1037/a0026609

Umemuro, H. & Yamashita, J. (2003). Detection of user's confusion and surprise based on pupil dilation. *The Japanese Journal of Ergonomics*, *39*(4), 153–161. doi: 10.5100/jje.39.153

van Dooren, M., de Vries, J. J. G. G. J. & Janssen, J. H. (2012). Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology and Behavior*, *106*(2), 298–304. doi: 10.1016/j.physbeh.2012.01.020

Van Gerven, P. W. M., Paas, F., Van Merriënboer, J. J. G. & Schmidt, H. G. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology*, *41*(2), 167–174. doi: 10.1111/j.1469-8986.2003.00148.x

van Gog, T. & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist*, *43*(1), 16–26. doi: 10.1080/00461520701756248

van Gog, T. & Scheiter, K. (2010). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction*, *20*(2), 95–99. doi: 10.1016/j.learninstruc.2009.02.009

van Marlen, T., van Wermeskerken, M., Jarodzka, H. & van Gog, T. (2016). Showing a model's eye movements in examples does not improve learning of problem-solving tasks. *Computers in Human Behavior*, *65*, 448–459. doi: 10.1016/j.chb.2016.08.041

van Merriënboer, J. J. G. & Sweller, J. (2005). Cognitive Load Theory and Complex Learning: Recent Developments and Future Directions. *Educational Psychology Review*, *17*(2), 147–177. doi: 10.1007/s10648-005-3951-0

Van Waes, L., Leijten, M. & Van Weijen, D. (2009). Keystroke logging in writing research: Observing writing process with inputlog. *German as a Foreign Language*, *2*(3), 41–64.

Vandewaetere, M., Desmet, P. & Clarebout, G. (2011). The contribution of learner characteristics in the development of computer-based adaptive learning environments. *Computers in Human Behavior*, *27*(1), 118–130. doi: 10.1016/j.chb.2010.07.038

Veenman, M. V. J. (2005). The assessment of metacognitive skills: What can be learned from multimethod designs? In C. Artelt & B. Moschner (Eds.), *Lernstrategien und Metakognition: Implikationen für Forschung und Praxis* (pp. 75–97). Berlin: Waxmann Verlag.

Veenman, M. V. J. (2011). Learning to Self-Monitor and Self-Regulate. In P. Alexander & R. E. Mayer (Eds.), *Handbook of Research on Learning and Instruction*. New York: Routledge.

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M. & Afflerbach, P. (2006). Metacognition and learning: conceptual and methodological considerations. *Metacognition and Learning*, *1*(1), 3–14. doi: 10.1007/s11409-006-6893-0

Veletsianos, G. & Russell, G. S. (2014). Pedagogical Agents. In J. M. Spector, M. D. Merrill, J. Elen & M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 759–769). New York, NY: Springer New York. doi: 10.1007/978-1-4614-3185-5_61

Verbert, K., Duval, E., Klerkx, J., Govaerts, S. & Santos, J. L. (2013). Learning Analytics Dashboard Applications. *American Behavioral Scientist*, *57*(10), 1500–1509. doi: 10.1177/0002764213479363

Veronikas, S. & Maushak, N. (2005). Effectiveness of Audio on Screen Captures in Software Application Instruction. *Journal of Educational Multimedia and Hypermedia*, *14*(2), 199–205.

Vicente, A. De & Pain, H. (2002). *Intelligent Tutoring Systems*. (S. A. Cerri, G. Gouardères & F. Paraguaçu, Eds.)*Intelligent tutoring systems* (Vol. 2363). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/3-540-47987-2

Vicente, A. De & Pain, H. (2003). Validating the Detection of a Student ' s Motivational State. In *Proceedings of the Second International Conference on Multimedia Information & Communication Technologies in Education* (pp. 2004–2008).

Viscomi, R., Davies, A. & Duran, M. (2015). *Using WebPageTest : web performance testing for novices and power users*. Sebastopol: O'Reilly.

Vizer, L. M., Zhou, L. & Sears, A. (2009). Automated stress detection using keystroke and linguistic features: An exploratory study. *International Journal of Human Computer Studies*, *67*(10), 870–886. doi: 10.1016/j.ijhcs.2009.07.005

Vollmeyer, R. & Rheinberg, F. (1998). Motivationale Einflüsse auf Erwerb und Anwendung von Wissen in einem computersimulierten System. *Zeitschrift Für Pädagogische Psychologie*, *12*(1), 11–23.

W3C World Wide Web Consortium. (1995). The Common Log File Format. Retrieved March 14, 2018, from https://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format

W3C World Wide Web Consortium. (2005). Document Object Model. Retrieved February 10, 2017, from https://www.w3.org/DOM/

Walters, G. J. (2001). *Human Rights in an Information Age*. Toronto: University of Toronto Press.

Wang, H., Li, Y., Hu, X., Yang, Y., Meng, Z. & Chang, K. M. (2013). Using EEG to improve massive open online courses feedback interaction. *CEUR Workshop Proceedings*, *1009*, 59–66.

Ware, C. (2013). *Information Visualization - Perception for Design*. New York: Morgan Kaufmann Publishers. doi: 10.2307/20206579

Watson, D., Clark, L. a & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–70. doi: 10.1037/0022-3514.54.6.1063

WebPageTest.org. (2018). Metrics - WebPagetest Documentation. Retrieved March 19, 2018, from https://sites.google.com/a/webpagetest.org/docs/using-webpagetest/metrics

Wengelin, A. (2006). Examining pauses in writing: Theories, methods and empirical data. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer key-stroke logging and writing: Methods and applications* (pp. 107–130). Oxford, UK: Elsevier.

White, G. L. & Sivitanides, M. P. (2009). A Theory of the Relationships between Cognitive Requirements of Computer Programming Languages and Programmers' Cognitive Characteristics. *Journal of Information Systems Education*, *13*(1), 59–66.

Wichmann, A. & Leutner, D. (2009). Inquiry Learning. *Zeitschrift Für Pädagogische Psychologie*, *23*(2), 117–127. doi: 10.1024/1010-0652.23.2.117

Wierwille, W. W. & Eggemeier, F. T. (1993). Recommendations for Mental Workload Measurement in a Test and Evaluation Environment. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *35*(2), 263–281. doi: 10.1177/001872089303500205

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-Regulated Learning and Academic Achievement* (pp. 153–190). New York, NY: Lawrence Erlbaum Associates.

Winne, P. H. & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker & J. Dunlosky (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Mahwah: Lawrence Erlbaum Associates Publishers.

Winne, P. H. & Hadwin, A. F. (2013). nStudy: Tracing and Supporting Self-Regulated Learning in the Internet. In R. Azevedo & V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies* (Vol. 28, pp. 293–308). New York: Springer Science+Business Media. doi: 10.1007/978-1-4419-5546-3_20

Winne, P. H., Jamieson-Noel, D. & Muis, K. (2002). Methodological issues and advances in researching tactics, strategies, and self-regulated learning. In P. R. Pintrich & W. Maehr (Eds.), *New Directions in Measures and Methods* (pp. 121–155). Stamford: JAI Press.

Winne, P. H. & Nesbit, J. C. (2009). Supporting Self-Regulated Learning with Cognitive Tools. In D. J. Hacker, J. Dunlosky & A. C. Graesser (Eds.), *The educational psychology series. Handbook of metacognition in education* (pp. 259–277). New York: Routledge/Taylor & Francis Group.

Winne, P. H., Nesbit, J. C. & Popowich, F. (2017). nStudy: A System for Researching Information Problem Solving. *Technology, Knowledge and Learning*, *22*(3), 369–376.

doi: 10.1007/s10758-017-9327-y

Winne, P. H. & Perry, N. E. (2000). Measuring Self-Regulated Learning. In M. Boekaerts, P. R. Pintrich & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 531–566). London: Academic Press. doi: 10.1016/B978-012109890-2/50045-7

Yamauchi, T., Seo, H., Choe, Y., Bowman, C. & Xiao, K. (2013). Assessing Emotions by Cursor Motions : An Affective Computing Approach. In *37th Annual Conference of the Cognitive Science Society* (pp. 2721–2726).

Yang, X., Zhang, L. & Yu, S. (2017). Can Short Answers to Open Response Questions Be Auto-Graded Without a Grading Rubric? In E. André, R. Baker, X. Hu, M. Mercedes, T. Rodrigo & B. du Boulay (Eds.), *Artificial Intelligence in Education, 18th International Conference - AIED 2017* (pp. 594–597). New York: Springer. doi: 10.1007/978-3-319-61425-0_72

Yannakakis, G. N., Hallam, J. & Lund, H. H. (2008). Entertainment capture through heart rate activity in physical interactive playgrounds. *User Modeling and User-Adapted Interaction*, *18*(1–2), 207–243. doi: 10.1007/s11257-007-9036-7

Yew, E. H. J. & Schmidt, H. G. (2012). What students learn in problem-based learning: a process analysis. *Instructional Science*, *40*(2), 371–395. doi: 10.1007/s11251-011-9181-6

Young, S. G. & Bowman, A. W. (1995). Nonparametric analysis of covariance. *Biometrics*, *51*, 920–931.

Zhang, M. & Quintana, C. (2012). Scaffolding strategies for supporting middle school students' online inquiry processes. *Computers and Education*, *58*(1), 181–196. doi: 10.1016/j.compedu.2011.07.016

Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary Educational Psychology*, *11*(4), 307–313. doi: 10.1016/0361-476X(86)90027-5

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, *81*(3), 329–339. doi: 10.1037/0022-0663.81.3.329

Zimmerman, B. J. (2000). Attaining Self-Regulation. In M. . Boekaerts, P. R. Pintrich & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 13–39). London: Academic Press. doi: 10.1016/B978-012109890-2/50031-7

Zimmerman, B. J. (2001). The Last Word. *Journal of Business Strategy*, *22*(3), 48–48. doi: 10.1108/eb040174

Zimmerman, B. J. (2008). Investigating Self-Regulation and Motivation: Historical Background, Methodological Developments, and Future Prospects. *American Educational Research Journal*, *45*(1), 166–183. doi: 10.3102/0002831207312909

Zimmerman, B. J. (2013). From Cognitive Modeling to Self-Regulation: A Social Cognitive Career Path. *Educational Psychologist*, *48*(3), 135–147. doi: 10.1080/00461520.2013.794676

Zimmerman, B. J. & Moylan, A. R. (2009). Self-regulation: where metacognition and motivation intersect. In D. J. Hacker, J. Dunlosky & A. C. Graesser (Eds.), *Handbook of Metacognition in Education* (pp. 299–315). New York, NY: Routledge.

Zimmermann, P. G., Gomez, P., Danuser, B. & Schär, S. G. (2006). Extending usability: putting affect into the user-experience. In *Proceedings of NordiCHI'06* (pp. 27–32).

Zorrilla, M. E., Menasalvas, E., Marín, D., Mora, E. & Segovia, J. (2005). Web Usage Mining Project for Improving Web-Based Learning Sites. In R. D. Moreno, F. Pichler & A. Q. Arencibia (Eds.), *10th International Conference on Computer Aided Systems Theory* (pp. 205–210). New York: Springer. doi: 10.1007/11556985_26

List of Tables

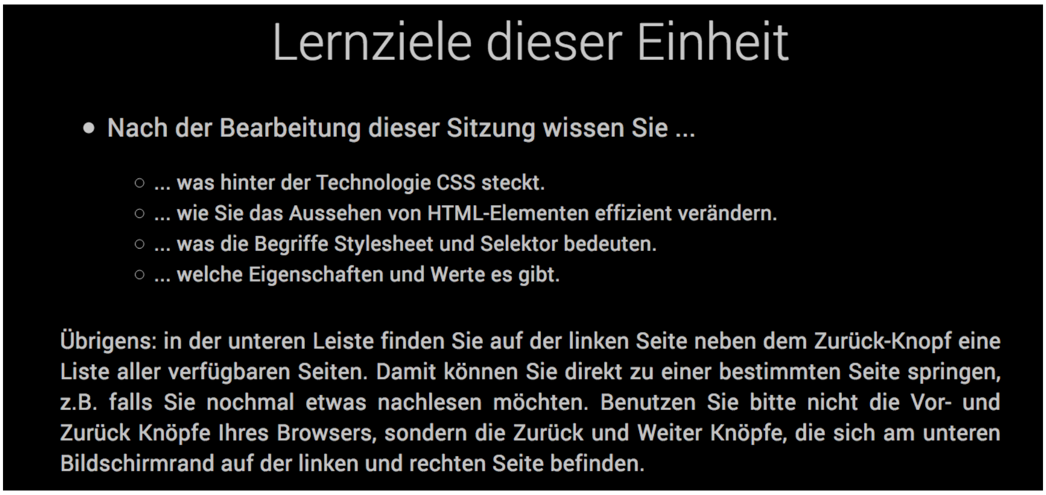
List of Figures

Appendix

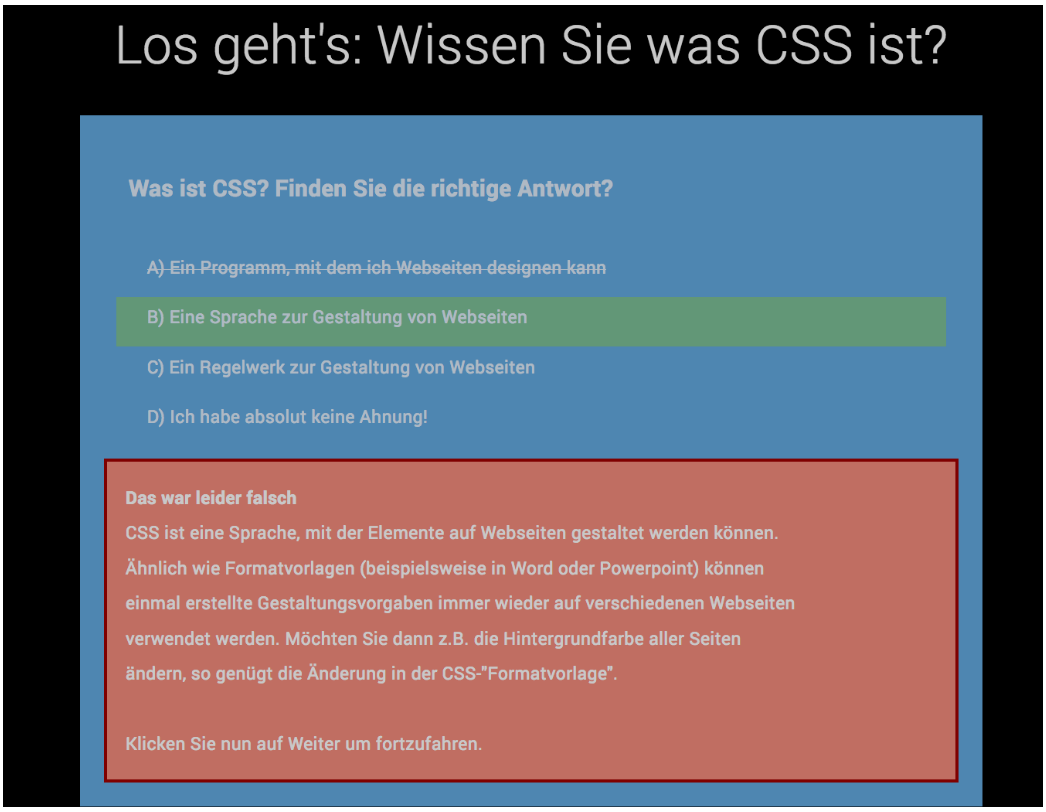
Appendix A – Learning Materials

Appendix A1: Learning Materials of Study 1 and 2: CSS

Learning content, page 1



Learning content, page 2:



Learning content, page 3:

Cascading Style Sheets

- CSS ist die Abkürzung für ***Cascading Style Sheets***
- Stylesheets funktionieren ähnliche wie Formatvorlagen in Word. Ein Stylesheet kann für viele verschiedene Webseiten eingesetzt werden
- CSS bietet vielfältige Möglichkeiten um Eigenschaften wie z.B. Positionen, Größen oder Farben von HTML-Elemente einfach zu verändern
- weil CSS das Aussehen von HTML-Elementen näher beschreibt, ergibt die Nutzung von CSS nur im Zusammenspiel mit HTML Sinn

Learning content, page 4:

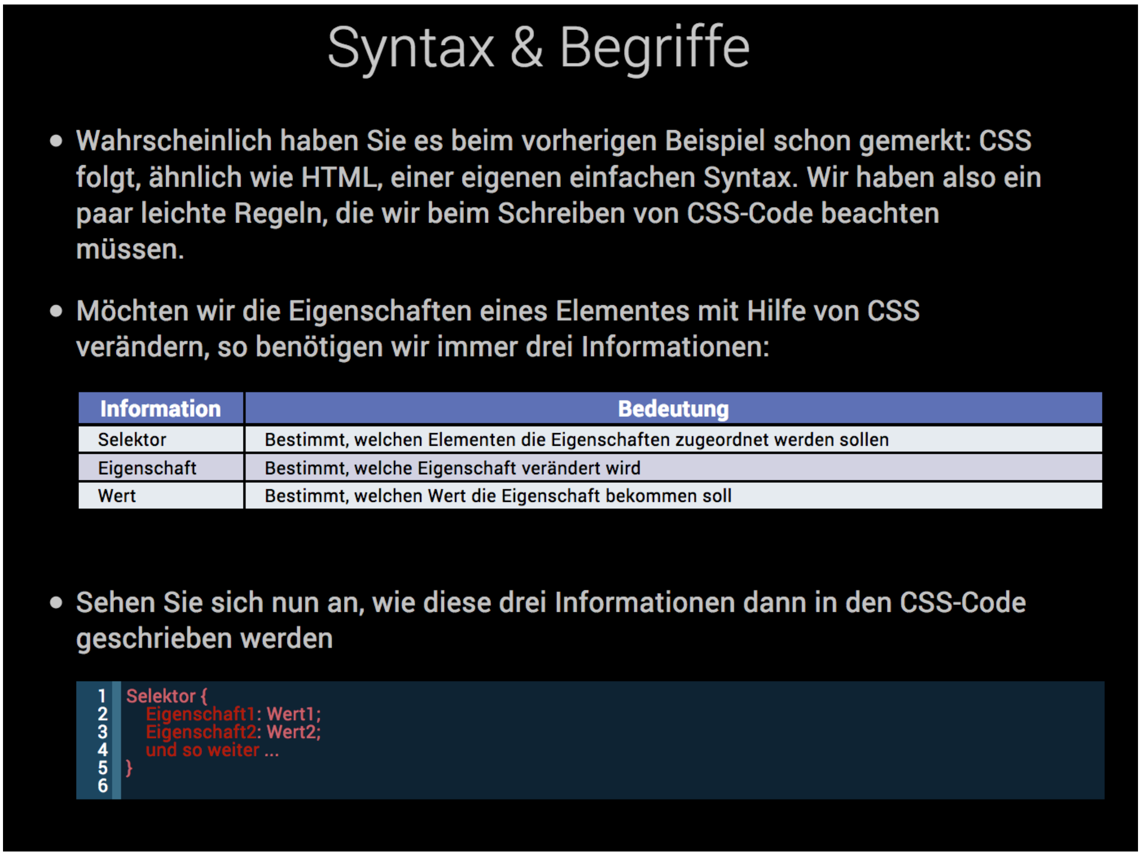
- Wie bei Formatvorlagen ist ein großer Vorteil von CSS, dass die einmal geschriebenen Styles einfach auf ganz verschiedenen HTML-Seiten eingebunden werden können.
- Eine Änderung der Farbe einer Überschrift muss also nur einmal vorgenommen werden und gilt anschließend für alle Überschriften der Webseite
- Unten sehen Sie einen Beispielcode für die Gestaltung einer Tabelle. Darunter sehen Sie die Tabelle.
- Sehen Sie sich den Code genau an und klicken Sie anschließend auf "Ausprobieren". Was passiert, wenn Sie den Style-Code aktivieren? Welche Code-Zeilen sind für welchen Teil des Aussehens verantwortlich?

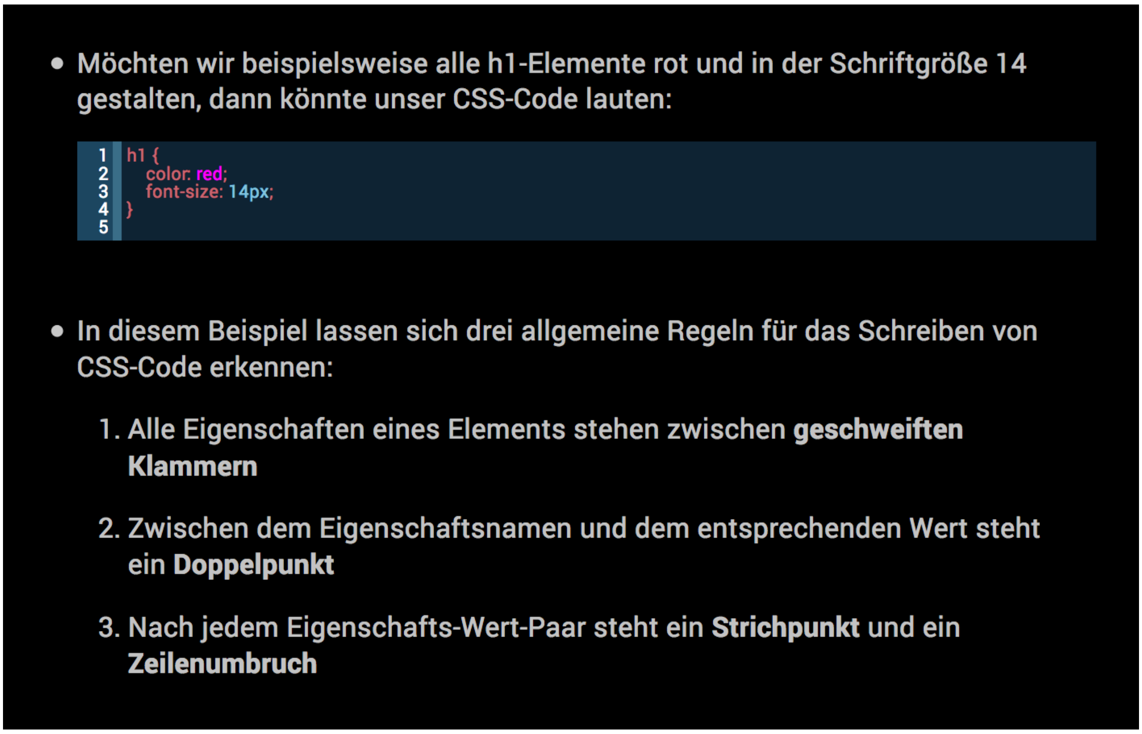
```
table {
   border: 10px dotted black;
   background-color:chocolate;
   color: blue;
   font-size: 20px;
   padding:20px;
}
```

Zelle 1	Zelle 2
Zelle 3	Zelle 4

Ausprobieren

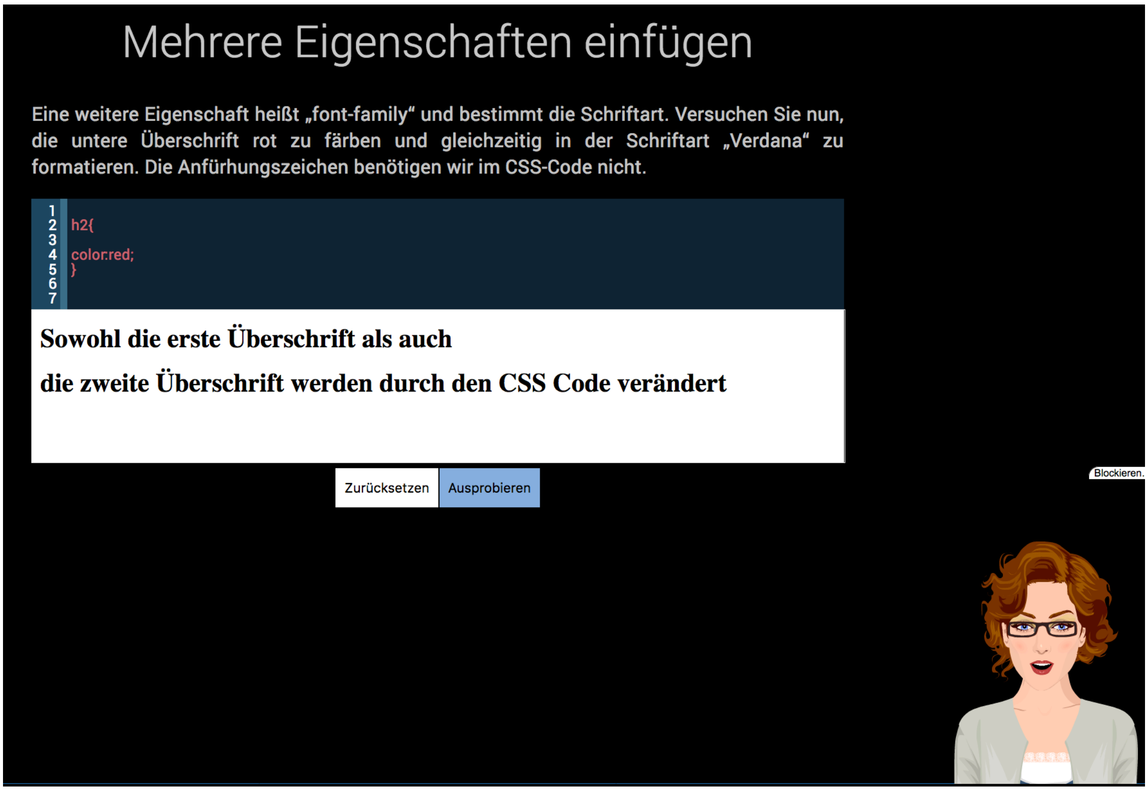
Learning content, page 5:





Learning content, page 6:

Learning content, page 7:



Learning content, page 8:

Verschiedene Werte von Eigenschaften

- Werte von Eigenschaften werden nicht immer in Worten angegeben, sondern sind manchmal auch Nummernwerte mit Einheiten.
- Ein Beispiel ist die Breite (***width***) und die Höhe (***height***) von Elementen. Diese Werte können beispielsweise in der Anzahl von Pixeln angegeben werden. Ähnlich wie ***cm*** für Centimeter werden Pixel mit ***px*** abgekürzt.
- Der Selektor für Bilder heißt wie das HTML-Element: ***img***
- Versuchen Sie einmal, das untenstehenden Bild auf die Breite 200px zu setzen.

```
1
2
3
4
```

Zurücksetzen | Ausprobieren

Learning content, page 9:

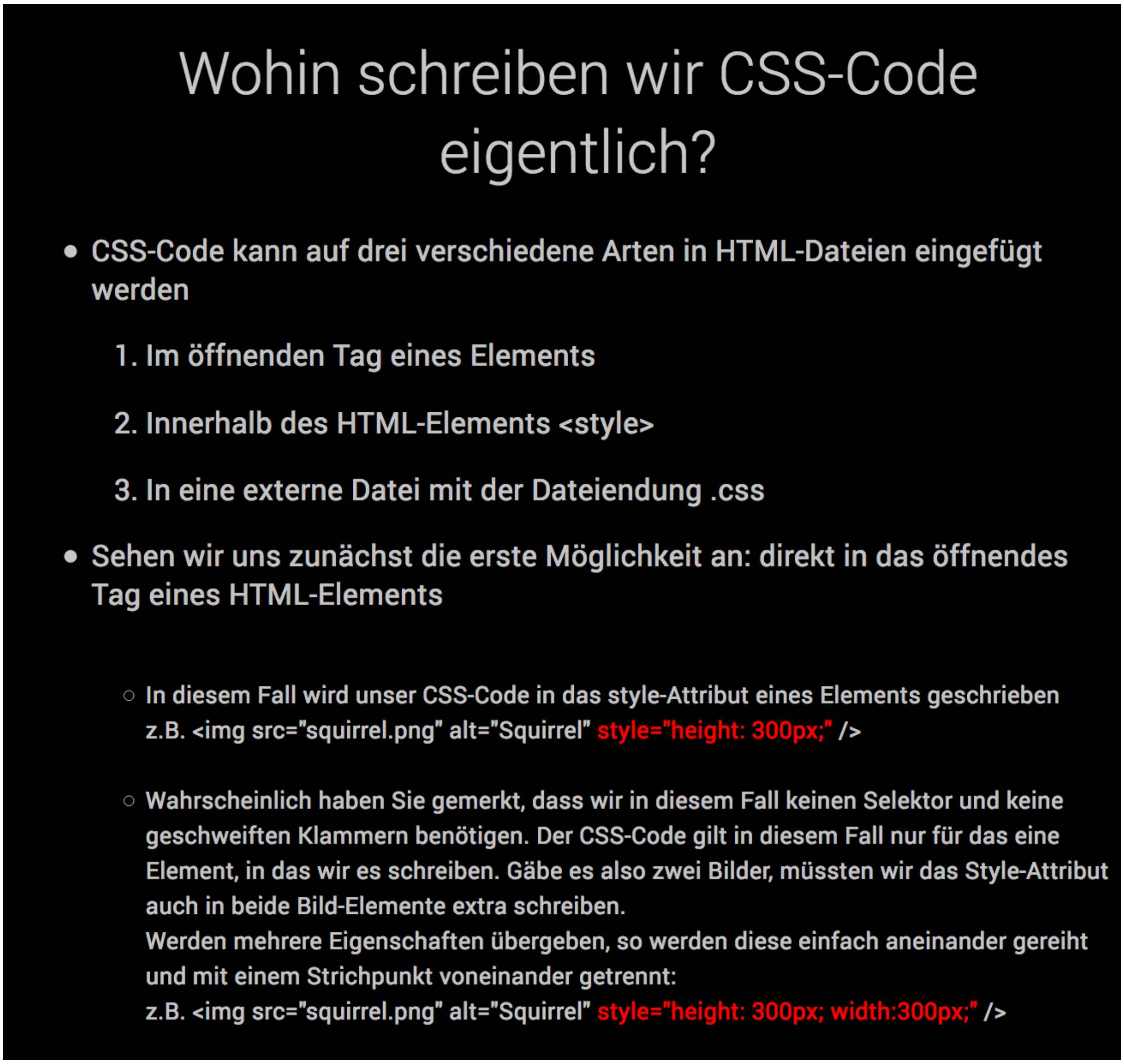
Weitere CSS-Eigenschaften

- In CSS gibt es sehr viele Eigenschaften, die wir nutzen können, um Elemente zu beschreiben
- Sehen Sie sich diese Liste mit Beispielen an. Können Sie sich einige davon bereits merken?

Eigenschaft	Bedeutung	Beispiel
background-color	Bestimmt die Hintergrundfarbe des Dokuments	background-color: chocolate;
Border	Bestimmt den Rahmen, z.B. von einer Tabelle	border: 1px;
text-align	Verändert die horizontale Ausrichtung von Elementen, z.B. links oder zentriert	text-align: left;
top / left	Bestimmt den Abstand von oben / links	top: 20px;
float	Bestimmt, ob und auf welche Seite Elemente "fließen", sich positionieren	float: right;
font-style	Verändert die Schrift zu kursiv	font-style: italic;
font-weight	Verändert die Schrift zu fett	font-weigth: bold;

- Diese und viele weitere Eigenschaften müssen Sie nicht alle auswendig im Kopf behalten. Eine vollständige Übersicht über alle CSS-Eigenschaften geben sog. CSS-Referenzen, in der Sie jederzeit nachschlagen können.

Learning content page 10:



Learning content, page 11:

Learning content, page 12:

Wohin schreiben wir CSS-Code eigentlich?

- Möglichkeit 3: in eine externe CSS-Datei
  - Wir können unsere CSS-Codes auch in eine extra Datei verpacken. Diese Datei muss die Endung .css tragen
  - Damit der Computer weiß, dass er diese Datei verwenden soll, müssen wir sie in der HTML-Datei einbinden. Das funktioniert sehr ähnlich wie bei einem Bild. Das Element dazu heißt allerdings „link“ und wird in den HEAD-Bereich einer HTML-Datei geschrieben
  - Sehen Sie sich im Beispielcode genau an, wo die externe Datei eingebunden wird:

```
<!DOCTYPE html>
<html>
  <head>
    <title>Webseite</title>
    <link rel="stylesheet" type="text/css" href="style.css" />
  </head>
<body>
<h1>Test</h1>
</body>
</html>

```

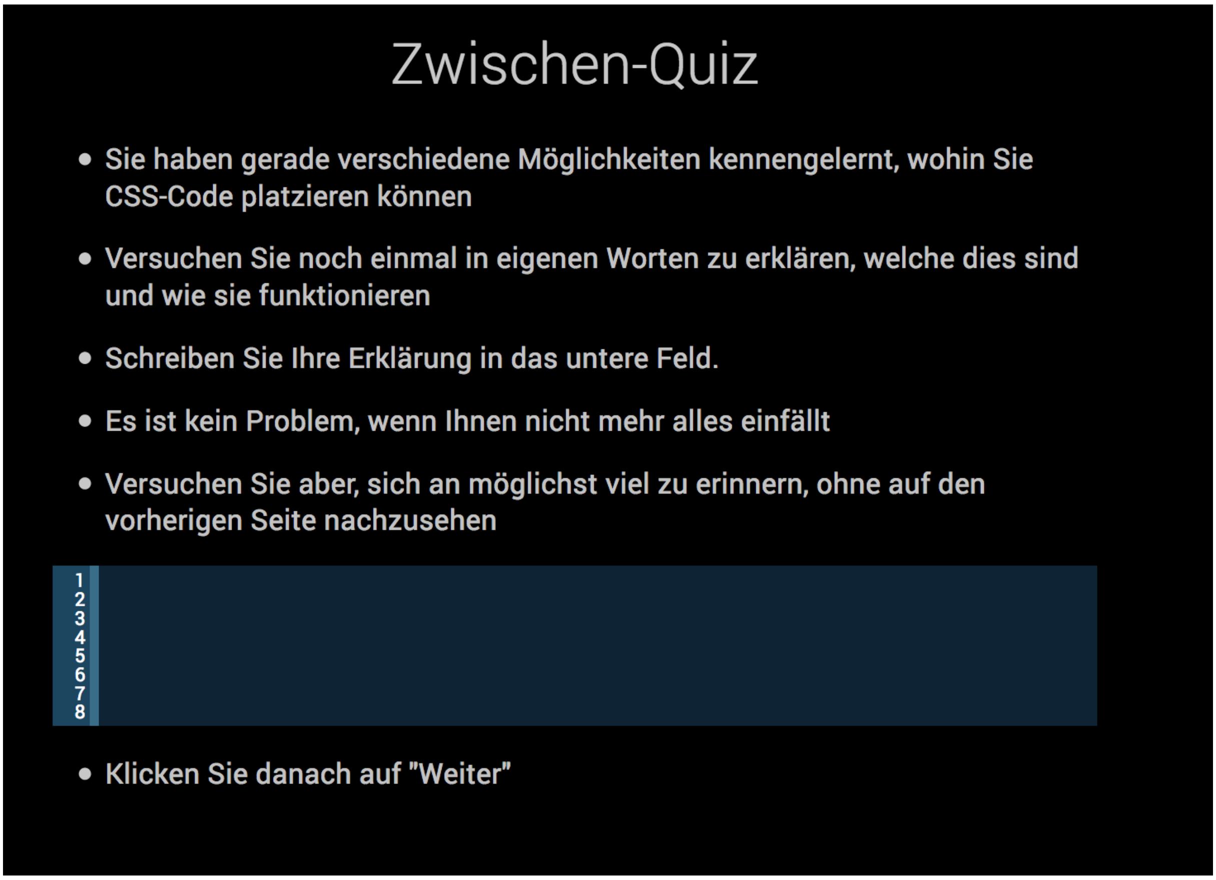
  - In der CSS-Datei style.css steht dann der entsprechende CSS-Code mit Selektoren, Eigenschaften und Werten:

```
h1{
    color:green;
    text-align:center;
}

```

- Ob wir den CSS-Code in eine externe Datei oder innerhalb der HTML-Datei schreiben, ist für das Aussehen egal. Der Vorteil von externen Dateien ist, dass wir sie einfach auf sehr vielen Webseiten einbauen können. Möchten wir dann beispielsweise die Hintergrundfarbe aller Webseiten ändern, so müssen wir die Änderung nur einmal in der externen CSS-Datei vornehmen.

Learning content page 13:



Learning content, page 14:

Welche Selektoren gibt es?

- Selektoren legen fest, welche HTML-Elemente von den veränderten Eigenschaften betroffen sind
- Bisher haben wir nur eine Art von Selektoren kennengelernt, wir haben z.B. h1 { ... } geschrieben, um alle Überschriften auf Ebene 1 zu verändern
- Um zu bestimmen, welchen Elementen die Eigenschaften zugeordnet werden sollen, gibt es aber drei verschiedene Typen von Selektoren.
  1. **Element-Selektoren** gelten für alle Elemente in der HTML-Datei. Diese haben wir bislang verwendet, Sie kennen diesen Typ von Selektor also bereits
  2. **Klassen-Selektoren** gelten für alle Elemente, deren HTML-Attribut "class" den selben Wert hat
  3. **ID-Selektoren** gelten nur für einzelne Elemente, z.B. nur für eine Überschrift von vielen auf einer Seite
- An einem Beispiel wird das klarer: unten sehen Sie sechs verschiedene Elemente. Drei große Überschriften (<h1>) und drei kleine Überschriften (<h2>)
- Möchten wir nun die Farbe aller Überschriften auf Ebene 1 in rot verändern, so können wir den unten folgenden Code schreiben

```
h1 {
  color: red;
}
```

Volleyball

Skateboarden

Golf

Surfen

Fußball

Yoga

- Die Element Selektoren wählen also nur die h1-Elemente aus und wenden unser Styling darauf an. Die kleinen Überschriften sind nicht betroffen.
- Sehen wir uns auf der nächsten Seite die Klassen-Selektoren an.

Learning content, page 15:

Die Klassen-Selektoren

- Bislang haben wir immer alle Vorkommnisse eines bestimmten Elements verändert, z.B. alle Überschriften auf Ebene 1
- Was aber, wenn wir in unserem Beispiel diejenigen Sportarten rot markieren möchten, bei denen einen Ball verwendet wird?
- Sehen Sie sich das Beispiel an unten an: sowohl in den großen Überschriften (<h1>) als auch in den kleinen Überschriften befinden sich Sportarten mit Bällen (Volleyball, Fußball, Golf)
- Für diesen Fall können wir den sog. Klassen-Selektor verwenden.
- Dafür müssen wir dem Computer zunächst sagen, welche Überschriften zu einer Klassen gehören. Das machen wir mit dem HTML-Attribut „class“. Sehen Sie sich den folgenden Ausschnitt aus einer HTML-Datei an:

```html
<h1 class="ball">Volleyball</h1>
<h1>Skateboarden</h1>
<h1 class="ball">Golf</h1>
<h2>Surfen</h2>
<h2 class="ball">Fußball</h2>
<h2>Yoga</h2>
```

- Alle Sportarten mit Ball haben das HTML-Attribut "class" und den zugehörigen Wert "ball" bekommen
- Jetzt können wir diese "Klasse" im CSS-Code ansprechen, indem wir einen Punkt (.) vor den Klassennamen schreiben:

```css
.ball {
   color: red;
}
```

- Es werden also im Ergebnis nur diejenigen Überschriften rot gefärbt, die das Attribut "class" mit dem Wert "ball" besitzen

Volleyball

Skateboarden

Golf

Surfen

Fußball

Yoga

Learning page 16:

Die ID-Selektoren

- Möchten wir nur einzelne Elemente, z.B. nur eine Überschrift der Seite mit CSS gestalten, so können wir die sog. ID-Selektoren verwenden
- Dafür müssen wir zunächst ein Element mit dem ID-Attribut versehen.
- ID steht für Identifikation und ist ein eindeutiger Name. Eine ID darf deshalb immer nur einmal vorkommen, es dürfen also nicht zwei Elemente die gleiche ID haben.
- Im folgenden Beispiel hat jede Überschrift eine eigene, eindeutige ID

```
<h1 id="volleyball">Volleyball</h1>
<h1 id="skaten">Skateboarden</h1>
<h1 id="golf">Golf</h1>
<h2 id="surfen">Surfen</h2>
<h2 id="fussball">Fußball</h2>
<h2 id="yoga">Yoga</h2>

```

- Im CSS-Code können wir jetzt jede Sportart einzeln gestalten. Anstatt wie bei Klassen-Selektoren mit einem Punkt (.), greifen wir auf ID-Selektoren mit einem Hashtag(#) zu.

- Sehen Sie sich dafür diesen CSS-Beispielcode an:

```
#volleyball{
   color:red;
}
#skaten{
   color:green;
}
#golf{
   color:blue;
}
#surfen{
   color:chocolate;
}
#fussball{
   color:orange;
}
#yoga{
   color:purple;
}

```

- Es werden also im Ergebnis alle Überschriften unterschiedlich gefärbt:

Volleyball

Skateboarden

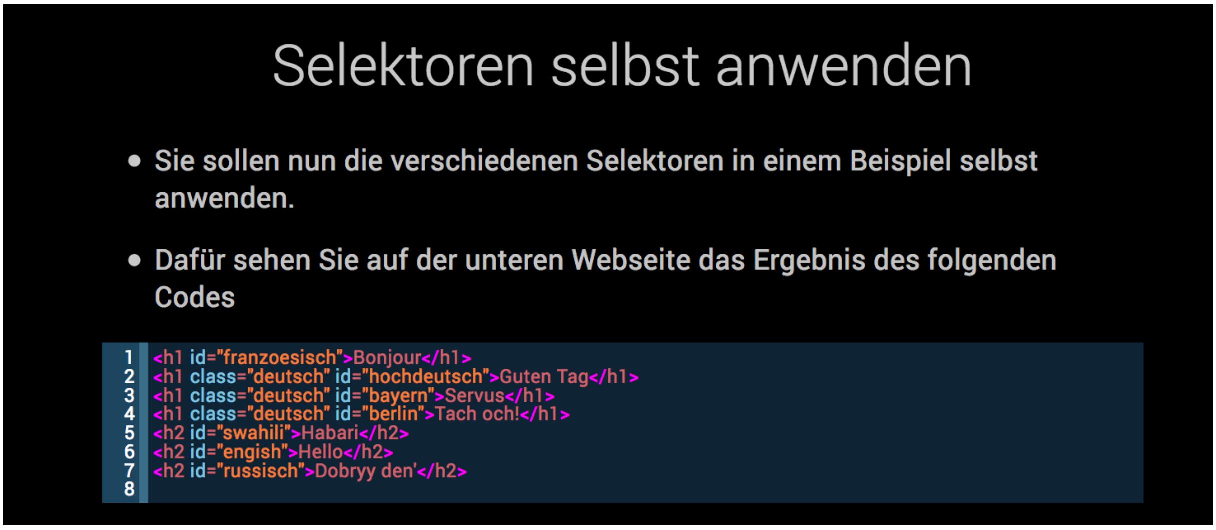
Golf

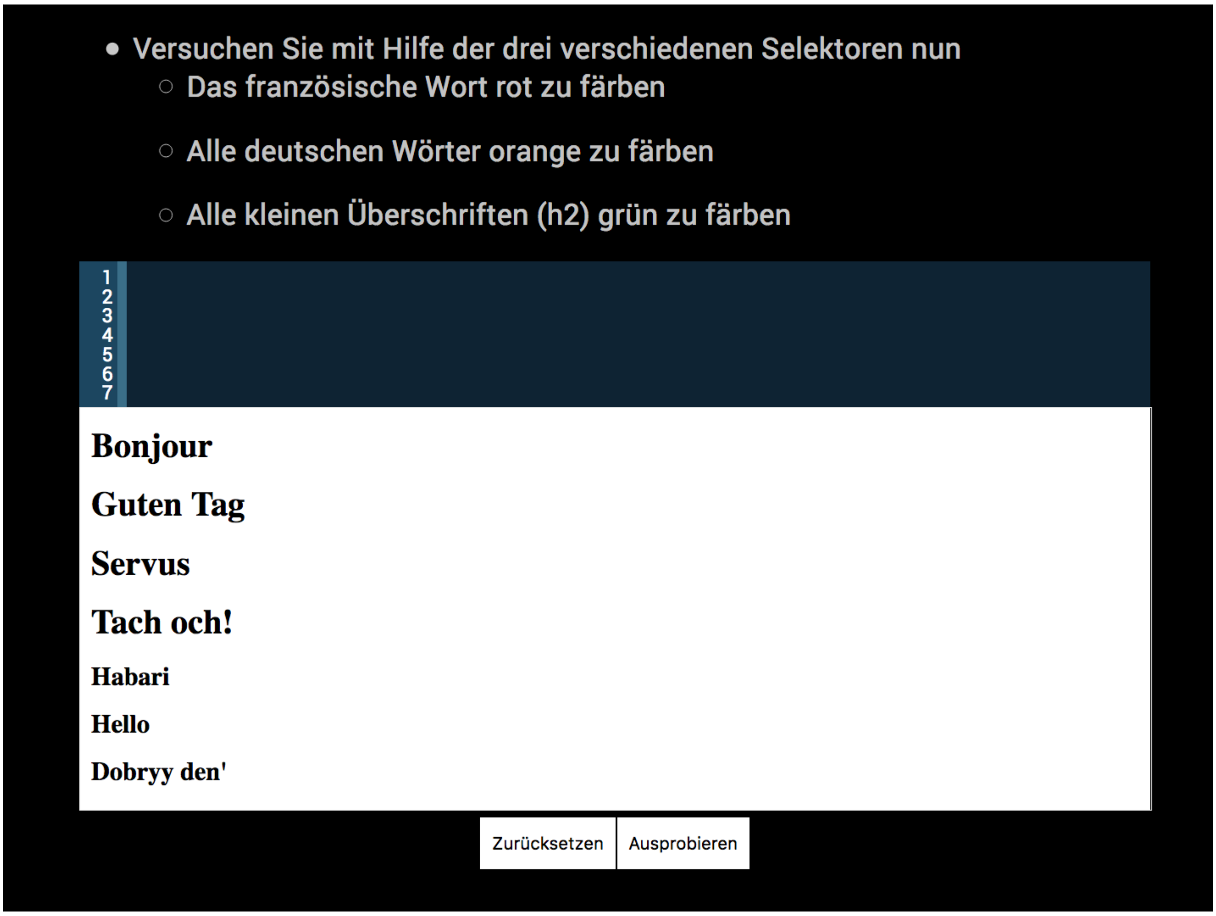
Surfen

Fußball

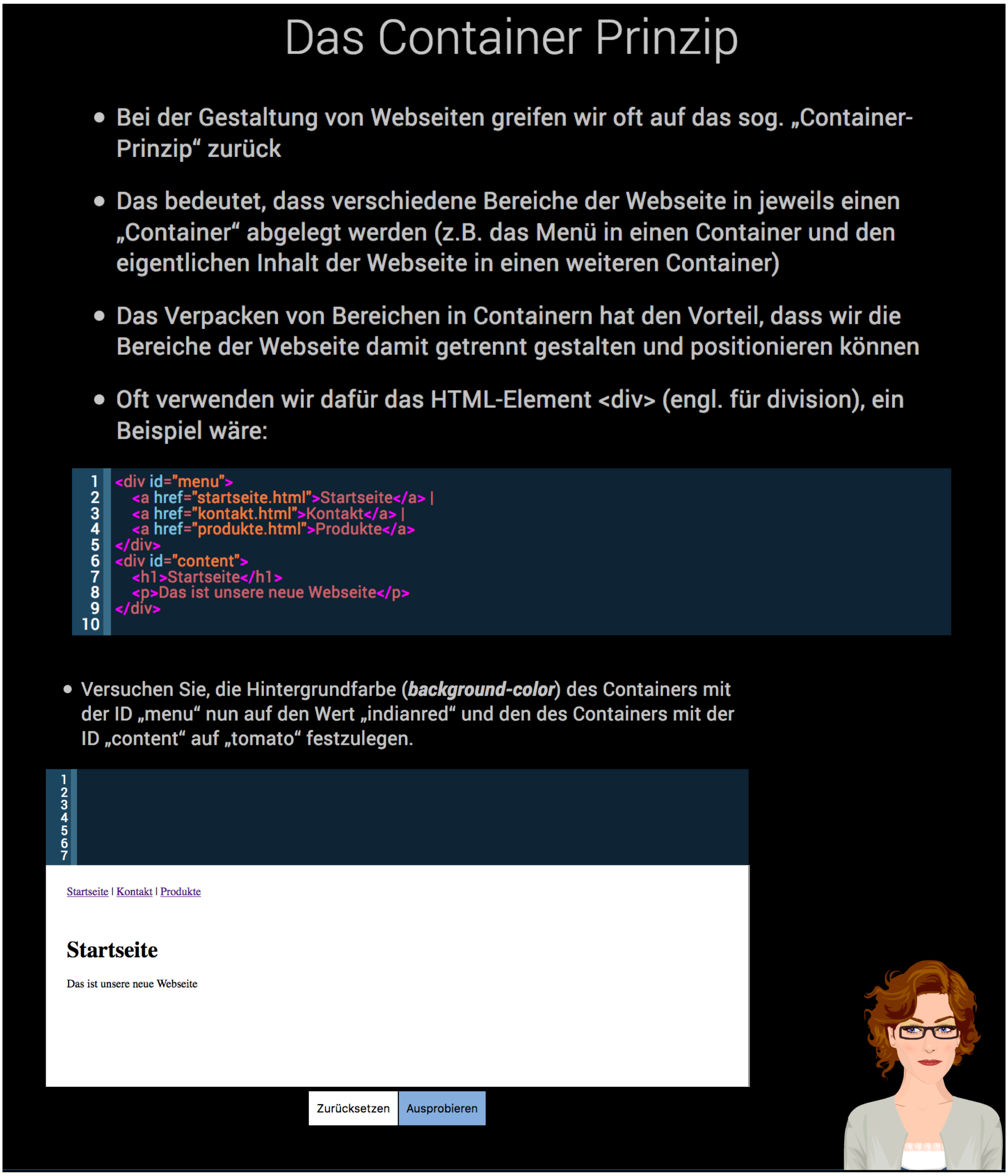
Yoga

Learning content, page 17:





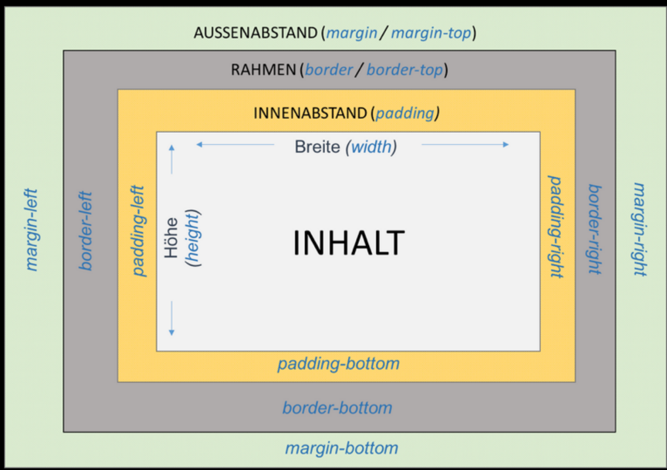
Learning content, page 18:



Learning content, page 19:

Das Box-Modell

- Mit dem Box-Modell können wir Abständen, Größe und Rahmen von HTML-Elementen verändern
- Sehen Sie sich das folgende Modell genau an.



- Das Modell besteht aus vier „Schichten“. Betrachten wir es schrittweise von außen nach innen.
  - Der „Außenabstand“ (margin) stellt den Abstand zu anderen Elementen der Webseite dar und wird in Pixeln angegeben. Er kann auch spezifisch für den Abstand nach oben (margin-top), unten (margin-bottom), rechts (margin-right) und links (margin-left) eingestellt werden.
  - Der „Rahmen“ kann in seiner Form, Farbe und Dicke gestaltet werden. Die Form kann beispielsweise eine durchgezogene oder gestrichelte Linien sein. Wir beschränken uns erst einmal auf die Dicke des Rahmens.
  - Der Abstand zwischen dem Rahmen und dem eigentlichen Inhalt wird „Innenabstand“ oder in CSS padding genannt. Auch dieser lässt sich generell (padding) bestimmen oder spezifisch für oben (padding-top), unten (padding-bottom), rechts (padding-right) und links (padding-left) festlegen.
  - Die Eigenschaften height und width legen die Höhe und Breite des Containers fest und werden auch in Pixeln angegeben.

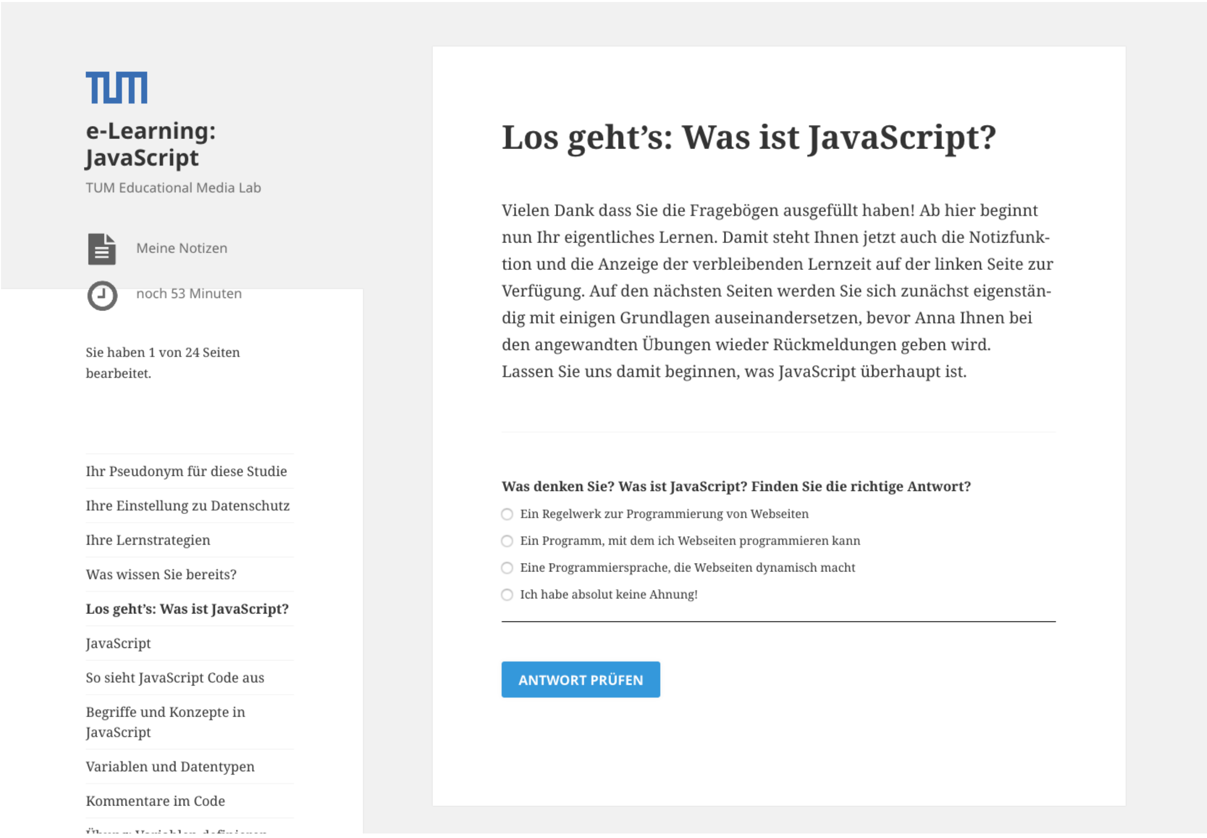
- Unten finden Sie ein Beispiel. Verändern Sie die Werte und sehen Sie sich an, was passiert.

border:	3px
margin:	20px
padding:	10px
padding-left:	0px
height:	100px
width:	300px

Learning content, page 20:

Appendix A2: Learning Materials of Study 4: JavaScript

Page 1:



Page 2:

e-Learning: JavaScript

TUM Educational Media Lab

Meine Notizen

noch 53 Minuten

Sie haben 2 von 24 Seiten bearbeitet.

JavaScript

- JavaScript ist eine sogenannte Programmiersprache.
- Neben JavaScript gibt es viele weitere Programmiersprachen, in denen Programme wie Word, Betriebssysteme wie Windows oder Spiele programmiert werden. Weit verbreitete andere Sprachen sind Python, Java und C++.
- JavaScript wird von Webbrowsern wie dem Internet Explorer, Firefox oder Google Chrome verstanden. Programme, die wir schreiben, können wir also direkt auf Webseiten im Internet ausführen, ohne ein zusätzliches Programm auf unserem Computer zu installieren. JavaScript Programme sind meistens Bestandteile von Webseiten.
- Die Grundlage von Webseiten bildet eine andere Sprache namens HTML (Hyper Text Markup Language). Mit dieser Sprache werden Texte, Bilder, Links, Tabellen usw. auf Webseiten gestaltet und formatiert. Wurde die Seite einmal fertig aus dem Internet geladen, verändert sich diese nicht mehr – sie ist also "statisch".
- JavaScript ist eine Zusatztechnik, die in Webseiten eingebaut werden kann um unsere Webseiten veränderbar, also "dynamisch" zu machen. Das bedeutet, dass wir Inhalte, Struktur und Aussehen einer Webseite verändern können, nachdem sie bereits im Webbrowser geladen wurde. JavaScript kümmert sich also zum Beispiel darum, dass Eingaben in Textfeldern überprüft werden und wir Rückmeldungen dazu bekommen. Oder dass wir bei Facebook flexibel zwischen Bildern hin- und herwechseln können, ohne jedesmal die gesamte Webseite neu zu laden. Nutzer können JavaScript in Ihren Browsern auch deaktivieren.
- Als Zusatztechnik zu HTML hat JavaScript auch Zugriff auf alle sog. HTML-Elemente (Bilder, Texte, Listen, Tabellen, etc.) der Webseite. Ein Programm in JavaScript könnte also z.B. den Text der Webseite einlesen, einige Wörter verändern, und ihn wieder zurück auf die Webseite schreiben.

WEITER ZU SO SIEHT JAVASCRIPT CODE AUS

Page 3:

e-Learning: JavaScript

TUM Educational Media Lab

Meine Notizen

noch 52 Minuten

Sie haben 3 von 24 Seiten bearbeitet.

So sieht JavaScript Code aus

- Code nennen wir die Aneinanderreihung und sinnvolle Verknüpfung von einzelnen Befehlen, die wir dem Computer mitteilen, damit er das tut, was wir möchten.
- Vielleicht haben Sie schon eine Vorstellung davon, wie Programmier-Code aussieht. JavaScript Code besteht immer aus aneinander gereihten Befehlen. Wir schreiben normalerweise jeden Befehl in eine neue Zeile – das macht den Code übersichtlicher. Jede Code-Zeile sollte mit einem Strichpunkt (;) enden. Der Computer liest dann eine Zeile nach der anderen ein und generiert unser Programm.
- Auf den ersten Blick wirkt Programmier-Code oft kryptisch – unten sehen Sie ein Beispiel. Keine Angst, an dieser Stelle müssen Sie den Code natürlich noch nicht verstehen. Die Zahlen auf der linken Seite dienen uns dabei nur zur Orientierung, in welcher Zeile wir uns befinden. Sie sind kein Bestandteil des eigentlichen Codes.
- Vielleicht können Sie sich ja bereits denken, was der Beispielcode tut? Mit dem Befehl "alert" wird ein kleines Fenster auf dem Bildschirm angezeigt, das die Information beinhaltet, die wir in den Klammern hinter "alert" schreiben.
- Sehen Sie sich den Code genau an und überlegen Sie, was passieren könnte. Führen Sie ihn dann mit einem Klick auf „Ausprobieren" aus und sehen Sie, ob Ihre Vermutung stimmt.

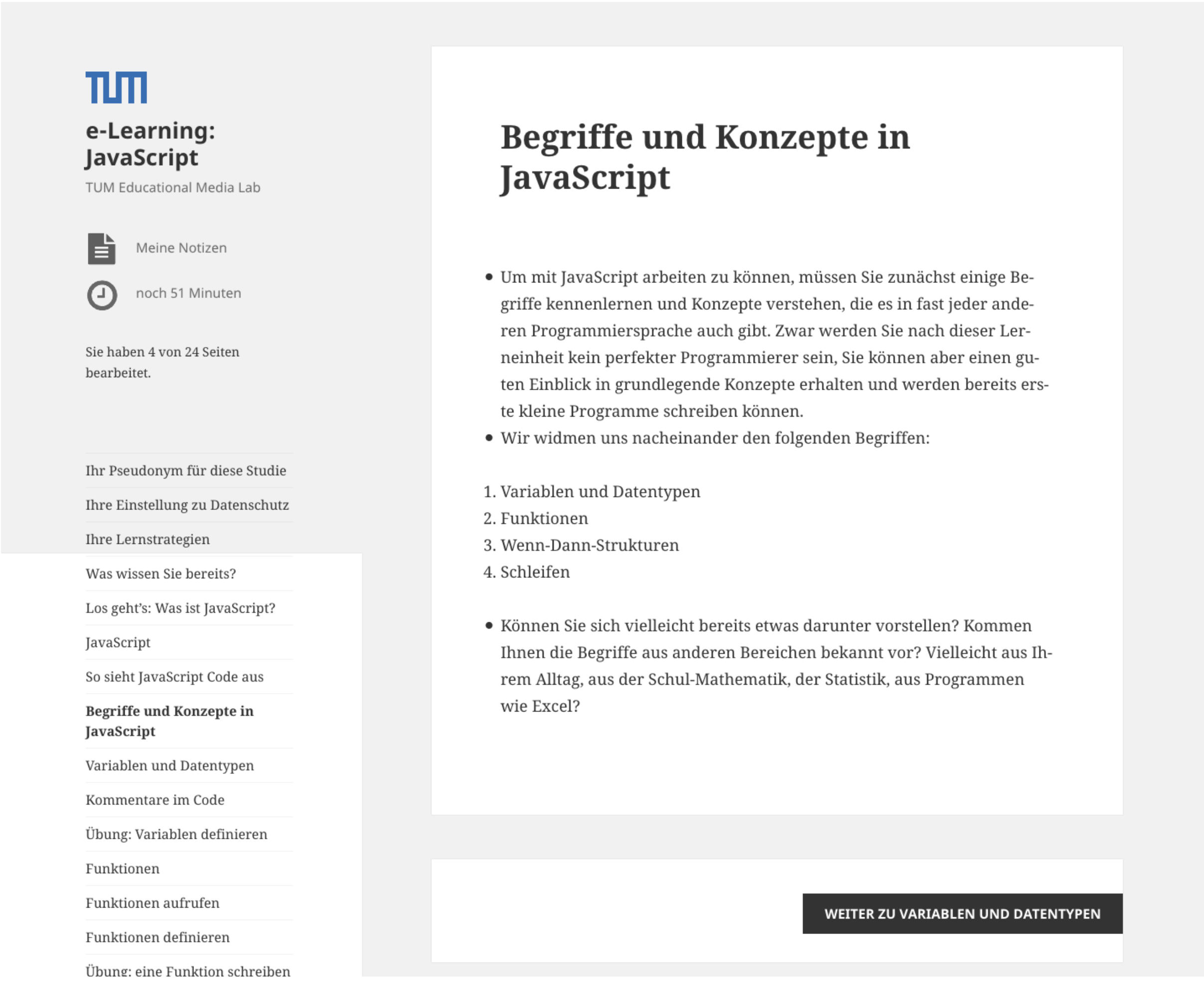
```
1 var a = 10;
2 var b = 19;
3 var ergebnis = a + b;
4 alert(ergebnis);
```

AUSPROBIEREN

- Bei Programmier-Code ist es wichtig zu unterscheiden, ob es sich um Befehle handelt, deren Namen in einer Sprache festgelegt sind oder ob es Werte sind, die wir selbst festlegen und benennen. So sind beispielsweise die Wörter *var* und *alert* vorher festgelegte Befehle und nicht veränderbar, während wir die Namen für *a*, *b* und *ergebnis* selbst und nach Belieben wählen können. Wir müssen beim Programmieren auch auf Groß-/Kleinschreibung achten, z.B. versteht der Computer unter *Alert* etwas anderes als unter *alert*. Zu all dem aber gleich noch mehr, klicken Sie auf *Weiter* um fortzufahren.

WEITER ZU BEGRIFFE UND KONZEPTE IN JAVASCRIPT

Page 4:



e-Learning: JavaScript

TUM Educational Media Lab

Meine Notizen

noch 51 Minuten

Sie haben 4 von 24 Seiten bearbeitet.

Begriffe und Konzepte in JavaScript

- Um mit JavaScript arbeiten zu können, müssen Sie zunächst einige Begriffe kennenlernen und Konzepte verstehen, die es in fast jeder anderen Programmiersprache auch gibt. Zwar werden Sie nach dieser Lerneinheit kein perfekter Programmierer sein, Sie können aber einen guten Einblick in grundlegende Konzepte erhalten und werden bereits erste kleine Programme schreiben können.
- Wir widmen uns nacheinander den folgenden Begriffen:

1. Variablen und Datentypen
2. Funktionen
3. Wenn-Dann-Strukturen
4. Schleifen

- Können Sie sich vielleicht bereits etwas darunter vorstellen? Kommen Ihnen die Begriffe aus anderen Bereichen bekannt vor? Vielleicht aus Ihrem Alltag, aus der Schul-Mathematik, der Statistik, aus Programmen wie Excel?

WEITER ZU VARIABLEN UND DATENTYPEN

Page 5:

e-Learning: JavaScript

TUM Educational Media Lab

Meine Notizen

noch 49 Minuten

Sie haben 4 von 24 Seiten bearbeitet.

Variablen und Datentypen

- Wir beginnen mit den sogenannten Variablen, die Sie wahrscheinlich schon aus der Mathematik kennen.
- Wie in der Mathematik, können wir auch in JavaScript Variablen definieren. In Variablen können wir Informationen wie Zahlen, Wörter oder Zustände zwischenzeitlich speichern, damit wir sie danach weiter verarbeiten können. Das können Sie sich vorstellen wie das menschliche Kurzzeitgedächtnis. Möchten Sie zwei Zahlen addieren, müssen Sie beide im Kurzzeitgedächtnis behalten um eine Operation, nämlich die Addition, damit durchzuführen.
- Anders als in der Mathematik, wo wir Variablen meistens nur mit einem Buchstaben bezeichnen, sollten wir beim Programmieren versuchen, jeder Variable einen sinnvollen Namen zu geben. Dieser muss immer mit einem Groß- oder Kleinbuchstaben beginnen. Dem Computer ist der Name der Variablen egal, natürlich könnten wir also auch einen einzelnen Buchstaben als Variablenname verwenden. Bei komplexeren Programmen kann es dann aber schnell unübersichtlich werden und wir wissen später nicht mehr, um welche gespeicherte Information es sich in der Variable handelt.
- Wie bereits erwähnt: Variablennamen dürfen nie mit einer Zahl beginnen. Außerdem dürfen keine Leerzeichen, Umlaute oder sonstige Sonderzeichen darin vorkommen.
- Den Code, den wir benutzen um eine Variable mit einer Information zu befüllen haben Sie im ersten Beispielcode schon einmal gesehen. Er lautete z.B.

```
var ergebnis = 10;
```

- Der einleitende Befehl heißt "var", dann steht der frei wählbare Variablenname gefolgt von einem Ist-Gleich Zeichen. Auf der rechten Seite des Ist-Gleich Zeichens steht der Wert der Variable. Die Leerzeichen vor und nach dem Ist-Gleich Zeichen können dort stehen, damit der Code übersichtlicher bleibt – funktionieren würde der Code aber auch ohne. Möchten wir den Code zur Definition einer Variable allgemein formulieren, dann lautet er

```
var Variablenname = Inhalt;
```

- Der Inhalt von Variablen kann dabei sehr unterschiedlich sein. Wir können nicht nur Zahlen, sondern auch andere Arten von Daten speichern. Die verschiedenen Arten von Daten nennen wir Datentypen. Wir können also z.B. Wörter und Sätze, Zustände wie "wahr" oder "falsch" oder Listen mit Wörtern speichern. Sehen Sie sich die wichtigsten Datentypen und die entsprechenden Beispiele dazu einmal an:

Vielen Dank für die Teilnahme an dieser Testung!

Art der Information	Datentyp	Beispiel
Ganze Zahlen	integer	5
Kommazahlen	float	0.423
Zeichenfolgen	string	"Das ist ein Text"
Wahr oder Falsch	boolean	true / false
Objekte	object	z.B. eine Referenz auf ein Bild: var img = new Image(5,5);
Listen mehrerer Informationen	array	["apfel","birne","orange"]

- In den folgenden frei erfundenen Beispielen sehen Sie, wie die Variablendefinitionen als JavaScript Code für die obige Tabelle von verschiedenen Datentypen aussehen würde. Erinnern Sie sich, die Variablennamen sind frei wählbar und könnten auch anders heißen. Achten Sie darauf, wann wir Anführungszeichen verwenden und wann nicht!

```
var alter = 5;
var groesse = 1.81;
var name = "Peter Weinhuber";
var istStudent = true;
var faecher = Array("Bio","Chemie","Philosophie");
```

- Bei Zahlen und bei Wahr-Falsch Zuständen (true / false) dürfen keine Anführungszeichen verwendet werden, bei Wörtern und Buchstaben müssen dagegen immer Anführungszeichen verwendet werden.
- Eine weitere Besonderheit ist, dass wir anstatt Kommata bei Zahlen einen Punkt verwenden. Wir müssen z.B 1.81 anstatt 1,81 schreiben!
- Die Typen Object und Array haben wir an dieser Stelle zwar vorgestellt, wir werden aber aus zeitlichen Gründen nicht mehr im Detail darauf eingehen.
- Wichtig ist noch zu wissen, dass der Inhalt einer Variable überschrieben wird, wenn wir den selben Variablennamen später noch einmal verwenden. Den Befehl *var* benötigen wir nur bei der ersten Definition der Variable.
- Welche Information wäre also in folgendem Beispiel-Code in der Variable "wetter" gespeichert? Überlegen Sie und klicken Sie im unteren Beispiel auf "Ausprobieren". Die letzte Zeile im Code ist dafür verantwortlich, dass sich ein kleines Fenster öffnet, in dem der Inhalt der Variable *wetter* angezeigt wird.

```
var wetter = "sonnig";
wetter = "regnerisch";
wetter = "wechselhaft";
alert(wetter);
```

AUSPROBIEREN

Page 6:

e-Learning: JavaScript

TUM Educational Media Lab

Meine Notizen

noch 49 Minuten

Sie haben 5 von 24 Seiten bearbeitet.

Kommentare im Code

- Bevor Sie nun Ihren ersten eigenen Programmier-Code schreiben, sehen wir uns noch kurz an, wie wir im Code den Überblick behalten.
- Innerhalb des Codes können wir sogenannte Kommentare einfügen. Kommentare können in den Code geschrieben werden, ohne dass der Computer sie interpretiert, also verarbeitet. Sie dienen nur den Programmierenden, um einen besseren Überblick über den Programm-Code zu haben.
- Kommentare werden mit zwei Schrägstrichen eingeleitet. Alles was in derselben Zeile dahinter steht, wird vom Computer nicht beachtet.
- Übrigens: möglicherweise wird Ihnen der Code nicht ganz angezeigt, Sie können im Code-Feld dann nach rechts und links scrollen, um den gesamten Code anzuzeigen.

```
// Das ist ein Kommentar
// In der nächsten Zeile werden wir eine Variable definieren:
var name = "Peter"; // Kommentare können auch erst hinter einem Befehl in der:
// Dieser Code wird vom Computer nicht beachtet: var name = "Fritz";
// var name = "Claudia"; // Und dieser?
alert(name);
```

Sicherlich wissen Sie nun, was nun in der Variable "name" gespeichert ist? Klicken Sie dann auf "Ausprobieren", um zu sehen ob Sie richtig lagen.

AUSPROBIEREN

- Möchten wir Kommentare über mehrere Zeilen schreiben, so können wir jede einzelne Zeile mit // einleiten. Alternativ ist es auch möglich, den Beginn und das Ende eines Kommentars zu markieren. Dafür benutzen wir /* für den Beginn und */ für das Ende des Kommentars. Alle dazwischenliegenden Zeilen werden dann als Kommentar betrachtet.

```
var name="Peter";
// Ein einzeiliger Kommentar
var name="Claudia";
/* Ein mehrzeiliger Kommentar:
Wir müssen innerhalb dieses Kommentars
nicht jede Zeile mit // beginnen,
um einen Kommentar einzuleiten.
Das Ende des Kommentars
markieren wir dann mit */
var name="Franz";
// Noch ein einzeiliger Kommentar.
```

WEITER ZU ÜBUNG: VARIABLEN DEFINIEREN

Page 7:

e-Learning: JavaScript

TUM Educational Media Lab

Meine Notizen

noch 48 Minuten

Sie haben 6 von 24 Seiten bearbeitet.

Übung: Variablen definieren

- Nun sollen Sie Ihren ersten eigenen JavaScript Code schreiben.
- Unten sehen Sie eine Tabelle mit den Variablen und Inhalten, die Sie als JavaScript Code definieren sollen. Haben Sie die Variablen korrekt definiert und auf "Ausprobieren" geklickt, so wird die entsprechende Zeile der Tabelle grün. Versuchen Sie es!

Variablenname	Inhalt
vorname	Peter
alter	18
student	true
fach	Maschinenbau

- Code:

Es scheint als hätten Sie folgende Variable genutzt, die nicht definiert ist: Maschinenbau
Beachten Sie, dass Daten vom Typ *String* (z.B. Wörter oder Sätze) stets in Anführungszeichen stehen müssen! (ausblenden)

```
var vorname = "Peter";
var alter = 18;
var student = false;
var fach = Maschinenbau;
```

AUSPROBIEREN

WEITER ZU FUNKTIONEN

Page 8:

e-Learning: JavaScript
TUM Educational Media Lab

Meine Notizen

noch 47 Minuten

Sie haben 7 von 24 Seiten bearbeitet.

Funktionen

- Den Begriff "Funktion" kennen Sie wahrscheinlich schon aus dem Mathematikunterricht. Dort sehen Funktionen z.B. so aus: $f(x) = 2*x^2$
- Wir konnten der Funktion eine Variable x "übergeben". Die Funktion hat dann Operationen am Wert dieser Variable ausgeführt, sie zum Beispiel dupliziert und quadriert und das Ergebnis zurückgegeben.
- Auch in JavaScript gibt es sog. Funktionen. Sie helfen uns dabei, Programme zu organisieren und bereits geschriebenen Code an anderer Stelle nochmal zu verwenden.
- Es gibt in JavaScript bereits vorgefertigte Funktionen für viele Aufgaben, z.B. um eine Nachricht auf dem Bildschirm der Nutzer anzuzeigen, die Quadratwurzel zu ziehen, eine zufällige Zahl zu erstellen und vieles mehr. Wir können aber auch selbst eigene Funktionen entwickeln – dazu aber erst später mehr.
- Sehen wir uns zunächst an, wie wir bereits existierende Funktionen in JavaScript aufrufen. Nehmen wir an, Sie möchten die Nutzer Ihrer Webseite begrüßen und dafür eine Meldung anzeigen. JavaScript bietet dafür eine Funktion mit dem Namen „alert", die Ihnen sicher bereits vorher aufgefallen ist.
- Was wir beim Funktionsaufruf zwischen die Klammern schreiben, nennen wir "Argumente" der Funktion. Als Argument empfängt die Funktion "alert" die Nachricht, die angezeigt werden soll. Sie muss wie alle Zeichenfolgen (Strings) in JavaScript in Anführungszeichen stehen.
- Versuchen Sie nun, den untenstehenden Code so zu verändern, dass eine Begrüßung auf Ihrem Bildschirm angezeigt werden.

```
alert("Hallo!");
```

AUSPROBIEREN

WEITER ZU FUNKTIONEN AUFRUFEN

Page 9:

e-Learning: JavaScript

TUM Educational Media Lab

Meine Notizen

noch 46 Minuten

Sie haben 8 von 24 Seiten bearbeitet.

Funktionen aufrufen

- Sicher ist Ihnen bereits die Schreibweise zum Aufruf von Funktionen aufgefallen. Das ist Teil der sog. Syntax, also den Schreibregeln von JavaScript. Im vorherigen Beispiel haben wir geschrieben:

```
alert("Hier steht ein Text");
```

- Wie bereits erklärt, nennen wir die Information, die zwischen den Klammern stehen, die "Argumente" der Funktion.
- Formulieren wir dies allgemein, dann benötigen wir also immer
  - Den Namen der Funktion
  - Klammer auf
  - Argumente, die die Funktion verarbeitet, z.B. unseren Text
  - Klammer zu
  - Strichpunkt
- Zusammengefasst und verallgemeinert lautet die Syntax für einen Funktionsaufruf:

```
funktionsname(argument);
```

- Eine Funktion kann auch mehrere Argumente empfangen, die sie dann verarbeitet. Dann werden die Argumente einfach aneinander gehängt und mit Kommata voneinander getrennt, also könnte z.B. eine Funktion, die zwei Zahlen addiert wie folgt aufgerufen werden:

```
addiere(2,15);
```

- Das erste Argument ist also die Zahl 2, das zweite Argument ist die Zahl 15.

- Es sind auch Funktionen denkbar, die keine Argumente benötigen, weil sie keine Informationen verarbeiten, die wir an sie übergeben. Dann rufen wir die Funktion einfach mit leeren Klammern auf. Ein Beispiel ist die Funktion, um zufällige Zahlen zwischen 0 und 1 zu generieren. Sie heißt Math.random() und ist bereits in JavaScript verfügbar. Drücken Sie auf "Ausprobieren" um neue Zufallszahlen generieren und anzeigen zu lassen. Versuchen Sie es mehrmals. Den Code dafür sehen Sie im Folgenden:

```
var zufallszahl = Math.random();
alert(zufallszahl);
```

AUSPROBIEREN

- Beachten Sie, dass die Funktion *Math.random()* einen Wert (die generierte Zahl) zurück gibt, der in der Variable *zufallszahl* gespeichert wird. Der Name dieser Variable ist beliebig gewählt und könnte auch ganz anders heißen.
- Andere Funktionen (z.B. *alert*) geben dagegen keinen Wert zurück. Die Zuweisung zu einer Variable (z.B. *var ergebnis = alert("Hallo");*) wäre daher nicht sinnvoll und die Variable bliebe leer.

Page 10:

e-Learning: JavaScript

TUM Educational Media Lab

Meine Notizen

noch 44 Minuten

Sie haben 9 von 24 Seiten bearbeitet.

Funktionen definieren

- Wir können auch selbst Funktionen „erfinden". Der Sinn von selbst definierten Funktionen ist, dass wir den selben Code nur einmal schreiben müssen und danach auf diese Funktion zugreifen können. Die simple Funktion "addiere" auf der vorherigen Seite ist zum Beispiel eine Funktion, die noch nicht in JavaScript existierte. Wir müssen sie definieren, damit sie zum Aufruf zur Verfügung steht.
- Im folgenden sehen Sie den Code für diese Funktion. Sehen Sie sich den Code genau an, wir möchten nun Schritt für Schritt nachvollziehen, was dort geschrieben wurde. Können Sie sich denken, was passiert? Drücken Sie auf "Ausprobieren", um Ihre Vermutung zu überprüfen.

```
1 function addiere(zahl1, zahl2){
2     var summe = zahl1 + zahl2;
3     return summe;
4 }
5
6 var ergebnis = addiere(2,15);
7 alert(ergebnis);
8 ergebnis = addiere(159,123);
9 alert(ergebnis);
```

AUSPROBIEREN

- Sicherlich lagen Sie bereits richtig. Sehen wir uns den Code nun genauer an, um ihn besser zu verstehen:
- Wichtig ist zunächst zu verstehen, dass der Code aus zwei Teilen besteht. In Zeile 1 bis 4 wird die Funktion definiert. Erst dann weiß der Computer, was er bei einem Aufruf der Funktion "addiere" zu tun hat. In den unteren Zeilen 6 bis 9 rufen wir die Funktion dann auf und zeigen das Ergebnis der addiere-Funktion mit der alert-Funktion an. Die alert Funktion müssen wir nicht definieren, weil sie bereits in JavaScript existiert.
- Wie Sie eine Funktion aufrufen, haben Sie auf der vorherigen Seite bereits gelernt: *funktionsname(argument);*
- In der ersten Zeile schreiben wir das Wort „function". Dieser Befehl sagt dem Computer, dass wir jetzt eine neue Funktion definieren. Danach steht der Name der Funktion, nämlich „addiere". Den Namen wählen wir ganz nach Belieben – dem Computer ist egal, wie die Funktion heißt. Den Namen benötigen wir später, um die Funktion aufzurufen. Anstatt *addiere* hätte unsere selbst geschriebene Funktion z.B. auch *zusammenzaehlen*, *summe* oder *sabrina* heißen können.

```
1 function addiere(zahl1, zahl2){
```

- In den Klammern stehen dann unsere zwei Argumente: zahl1 und zahl2. Diese sind mit einem Komma getrennt. Damit definieren wir, wie viele Argumente, also Variablen, unsere Funktion bei einem späteren Aufruf benötigt und wie diese innerhalb unserer Funktion heißen. Auch hier ist der Name der Argumente wieder beliebig wählbar.

- Dann folgt eine öffnende geschweifte Klammer. Sie markiert den Beginn unserer Funktion. Zwei Zeilen darunter sehen Sie die dazugehörige schließende geschweifte Klammer. Alles was zwischen der öffnenden und der schließenden Klammer steht, gehört zu unserer Funktion und wird vom Computer ausgeführt, wenn wir die Funktion später aufrufen. Beachten Sie, dass nach einer öffnenden Klammer kein Strichpunkt steht – der Grund dafür ist, dass der Befehl noch nicht abgeschlossen ist. Wir könnten auch die gesamte Funktion in einer Zeile definieren, aber das wäre sehr unübersichtlich.
- Sehen wir uns den Code zwischen den Klammern an. In Zeile 2 und 3 steht Folgendes:

```
var summe = zahl1 + zahl2;
return summe;
```

- Wie Sie bereits wissen, definiert der Befehl “var” eine neue Variable. Im Beispiel definieren wir die Variable mit dem Namen “summe” und weisen ihm die Summe der beiden Variablen, zahl1 + zahl2 zu.
- In Zeile 3 steht der Befehl “return”, gefolgt von unserer neu definierten Variable “summe”. Das bedeutet, dass unsere Funktion die Variable “summe” als Ergebnis zurück gibt. Wie sich das “Zurückgeben” auswirkt, sehen wir dann in Zeile 6 bis 9.
- Bislang haben wir die Funktion lediglich definiert. Beim Ausführen des Programms würde noch nichts passieren. Erst mit den weiteren Zeilen darunter, rufen wir die Funktion auch auf.

Zeile 6 bis 9: die Funktion aufrufen

- In Zeile 6 definieren wir die Variable “ergebnis” und rufen bei der Zuweisung, also nach dem Ist-Gleich Zeichen, die Funktion addiere mit den Argumenten 2 und 15 auf. Weil wir der Funktion gesagt haben, dass sie “summe” zurückgeben (“return”) soll, wird unser Variable “ergebnis” mit dem Ergebnis der Funktion “addiere” gefüllt.
- Nun ist das Ergebnis von 2 + 15 in der Variable “ergebnis” gespeichert. In Zeile 7 wird diese Variable an die Funktion “alert” weitergegeben, die Sie bereits kennen. Anschließend führen wir dasselbe nocheinmal für die Zahlen 159 und 123 aus. Beachten Sie dass die Variable “ergebnis” in Zeile 8 überschrieben wird – das vorherige Ergebnis aus Zeile 6 befindet sich also nicht mehr im Speicher des Computers.
- Führen Sie das Skript mit einem Klick auf “Ausprobieren” aus, um zu sehen was passiert.

WEITER ZU ÜBUNG: EINE FUNKTION SCHREIBEN

Page 11:

e-Learning: JavaScript

TUM Educational Media Lab

Meine Notizen

noch 40 Minuten

Sie haben 10 von 24 Seiten bearbeitet.

Übung: eine Funktion schreiben

- Sie haben auf den vorherigen Seiten gesehen, wie eine Funktion definiert und aufgerufen wird.
- Nun sollen Sie selbst eine Funktion namens "multipliziere" schreiben, die zwei Zahlen als Argumente empfängt, diese multipliziert und das Er-

Bitte warten ...

Schließen

n für die Multiplikation ist in JavaScript

ießend auf und lassen Sie das Ergebnis mit der alert-Funktion anzeigen.
- Hier sehen Sie als Stütze noch einmal das schematische Grundgerüst einer Funktion.

```
function  Funktionsname (argumente){
   // Code der Funktion
   return Rueckgabe;
}
```

- Ihr Code:

Dem Befehl 'return' sollte ein Variablenname oder Wert folgen, den Sie zurück geben möchten. (ausblenden)

```
function multipliziere (x,y) {
c = x*y;
return c;
}
```

AUSPROBIEREN

WEITER ZU EINGABEN VON NUTZERN EMPFANGEN

Page 12:

e-Learning: JavaScript

TUM Educational Media Lab

Meine Notizen

noch 37 Minuten

Sie haben 11 von 24 Seiten bearbeitet.

Eingaben von Nutzern empfangen

- Sie haben nun gesehen, dass Funktionen ein Ergebnis zurückgeben können.
- Eine weitere Funktion von JavaScript lautet "prompt". Sie bittet den Nutzer, etwas einzugeben und gibt diese Eingabe zurück. Als Argument empfängt die Funktion die Frage, die angezeigt werden soll.
- Auf diese Weise können wir Eingaben in Variablen speichern. Können Sie sich bereits denken, was im folgenden Code passiert?

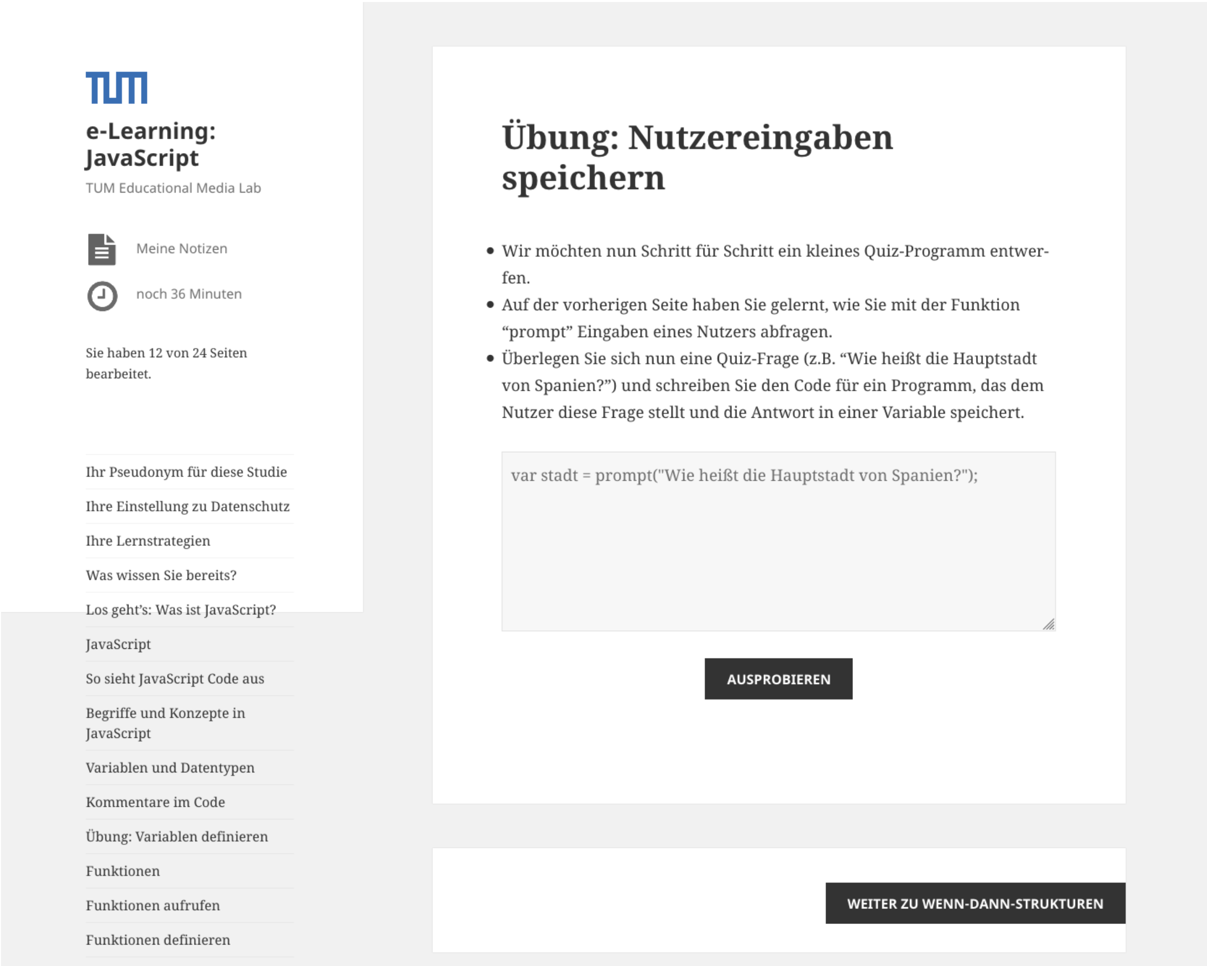
```
var eingabe = prompt("Wie lautet ihr Name?");
alert("Herzlich willkommen " + eingabe + "! Schön dass Sie hier sind!");
```

AUSPROBIEREN

- Die prompt-Funktion hat unsere Eingabe zurückgegeben und wir haben diese in der Variable "eingabe" gespeichert. Anschließend haben wir mit der alert-Funktion eine personalisierte Willkommens-Nachricht angezeigt.
- Sicherlich ist Ihnen aufgefallen, wie wir den Wilkommens-Text mit der Variable verknüpfen. Wie vorhin schon gesehen, müssen Texte immer in Anführungszeichen stehen. Möchten wir eine Variable in den Text einbauen, dann verknüpfen wir Text und Variable mit einem Plus-Zeichen. Weil danach wieder ein Text steht, schreiben wir nach der Variable "eingabe" nochmal ein Plus-Zeichen gefolgt vom Text in Anführungszeichen.

WEITER ZU ÜBUNG: NUTZEREINGABEN SPEICHERN

Page 13:


e-Learning: JavaScript
TUM Educational Media Lab
Meine Notizen
noch 36 Minuten
Sie haben 12 von 24 Seiten bearbeitet.
Ihr Pseudonym für diese Studie
Ihre Einstellung zu Datenschutz
Ihre Lernstrategien
Was wissen Sie bereits?
Los geht's: Was ist JavaScript?
JavaScript
So sieht JavaScript Code aus
Begriffe und Konzepte in JavaScript
Variablen und Datentypen
Kommentare im Code
Übung: Variablen definieren
Funktionen
Funktionen aufrufen
Funktionen definieren
Übung: Nutzereingaben speichern
• Wir möchten nun Schritt für Schritt ein kleines Quiz-Programm entwerfen.
• Auf der vorherigen Seite haben Sie gelernt, wie Sie mit der Funktion "prompt" Eingaben eines Nutzers abfragen.
• Überlegen Sie sich nun eine Quiz-Frage (z.B. "Wie heißt die Hauptstadt von Spanien?") und schreiben Sie den Code für ein Programm, das dem Nutzer diese Frage stellt und die Antwort in einer Variable speichert.
var stadt = prompt("Wie heißt die Hauptstadt von Spanien?");
AUSPROBIEREN
WEITER ZU WENN-DANN-STRUKTUREN

Page 14:

e-Learning: JavaScript

TUM Educational Media Lab

Meine Notizen

noch 36 Minuten

Sie haben 13 von 24 Seiten bearbeitet.

Wenn-Dann-Strukturen

- Häufig benötigen wir beim Programmieren sog. Wenn-Dann-Strukturen, um Programmabläufe zu steuern. Dabei sollen bestimmte Code-Teile nur ausgeführt werden, wenn eine bestimmte Bedingung zutrifft.
- Ein Beispiel könnte die Überprüfung der korrekten Antwort bei unserem Quiz sein. Ist die Antwort richtig, soll ein Lob angezeigt werden, ansonsten die richtige Antwort präsentiert werden.
- Um solche Strukturen umzusetzen, benötigen wir die Befehle *if* und *else*, also "falls" und "andernfalls".
- Die allgemeine Form beim Programmieren lautet dabei

```
if (Bedingung) {
    // Hier steht Code der ausgeführt wird,
    // wenn Bedingung zutrifft
}else{
    // Hier steht Code der ausgeführt wird,
    // wenn Bedingung NICHT zutrifft
    // Also in jedem anderen Fall
}
```

- Genau wie bei Funktionen steht der auszuführende Code immer zwischen öffnenden geschweiften Klammern und schließenden geschweiften Klammern.
- Wenn nur bei einer Bedingung etwas passieren soll, nicht aber in allen anderen Fällen, können wir den Teil ab "else" auch weglassen.
- Bedingungen, die in Klammern hinter dem Befehl "if" stehen, vergleichen dabei meist zwei Werte wie z.B. die Eingabe des Nutzers mit der richtigen Antwort. Zwischen den beiden Werten steht ein sog. "Vergleichsoperator".
- Aus Gründen der Übersichtlichkeit ist es hilfreich, wenn der Code zwischen den geschweiften Klammern etwas eingerückt ist. Das erreichen wir, indem wir beim Programmieren in der jeweiligen Zeile die Tabulator-Taste drücken (meist ist die Tabulator-Taste mit einem oder zwei Pfeilen beschriftet und befindet sich auf der linken Seite Ihrer Tastatur).
- Sehen Sie sich die folgenden Beispiele für solche Bedingungen und Vergleichsoperatoren an

Beispiel	Allgemeine Form	Vergleich-soperator	Ausformulier-te Form
antwort == "München"	A == B	==	Ist-Gleich
antwort != "München"	A != B	!=	Ist-Ungleich
alter > 18	A > B	>	Ist-Größer
alter >= 18	A >= B	>=	Ist-Größer-Oder-Gleich
alter < 18	A < B	<	Ist-Kleiner
alter <= 18	A <= B	<=	Ist-Kleiner-Oder-Gleich

- Bedingungen können also entweder wahr (true) oder falsch (false) sein. Um eine bessere Vorstellung von Bedingungen zu bekommen, sollen Sie nun einige Bedingungen im unteren Feld ein geben. Geben Sie z.B. 5 < 7 oder "Peter" != "Thomas" ein und sehen Sie was passiert. Verändern Sie dann die Zahlen und die Operatoren, um zu sehen, welche Bedingungen wahr oder falsch sind.

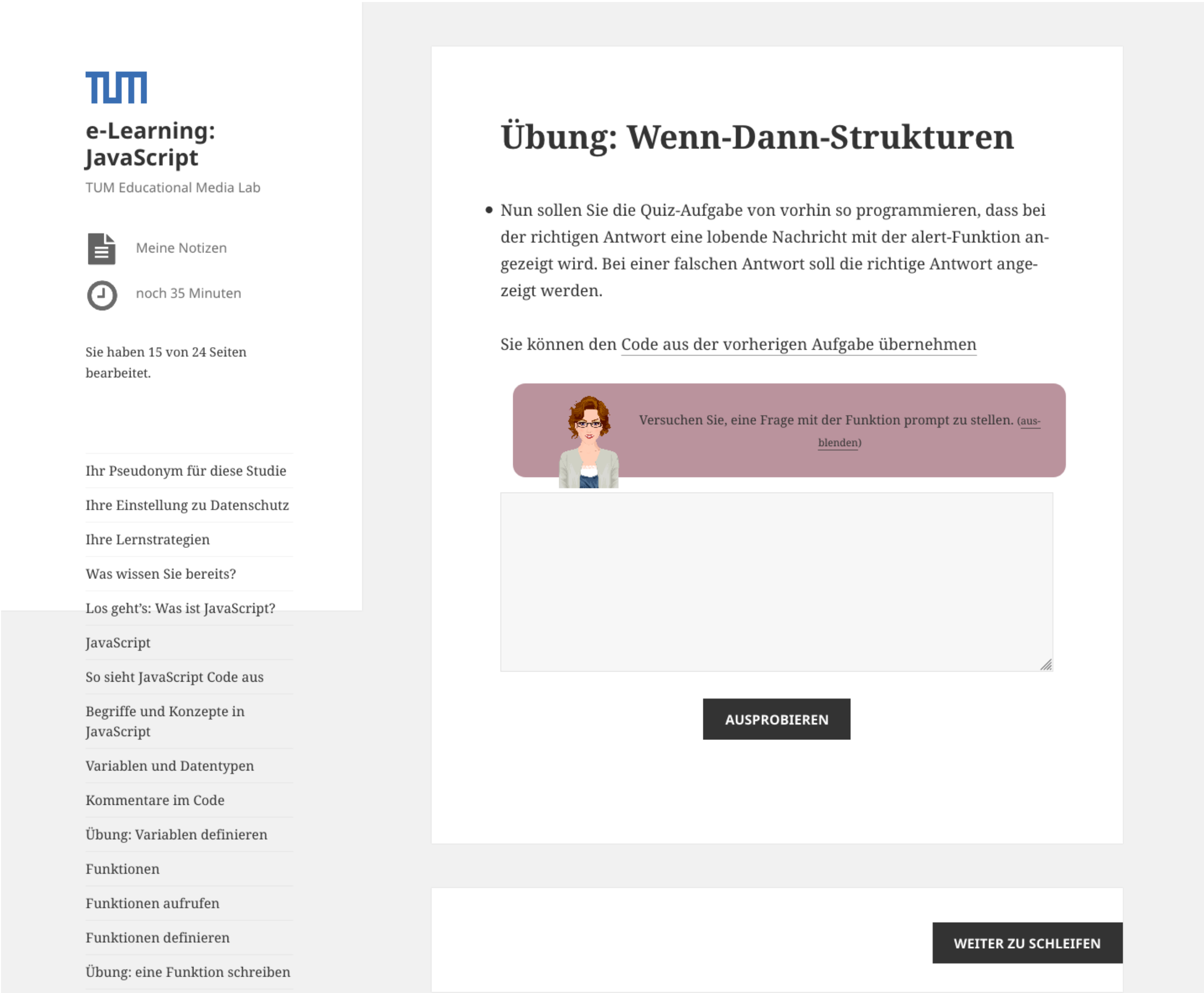
Bedingung: 2 == 3 Die Bedingung ist false / falsch

- Versuchen Sie nun, den untenstehenden Code so zu verändern, dass nur bei einem eingegebenen Alter über 18 die entsprechende Nachricht angezeigt wird.

```
var alter = prompt("Wie alt sind Sie?");
alert("Sie sind volljährig!");
```

AUSPROBIEREN

Page 15:


e-Learning: JavaScript
TUM Educational Media Lab
Meine Notizen
noch 35 Minuten
Sie haben 15 von 24 Seiten bearbeitet.
Ihr Pseudonym für diese Studie
Ihre Einstellung zu Datenschutz
Ihre Lernstrategien
Was wissen Sie bereits?
Los geht's: Was ist JavaScript?
JavaScript
So sieht JavaScript Code aus
Begriffe und Konzepte in JavaScript
Variablen und Datentypen
Kommentare im Code
Übung: Variablen definieren
Funktionen
Funktionen aufrufen
Funktionen definieren
Übung: eine Funktion schreiben
Übung: Wenn-Dann-Strukturen
Nun sollen Sie die Quiz-Aufgabe von vorhin so programmieren, dass bei der richtigen Antwort eine lobende Nachricht mit der alert-Funktion angezeigt wird. Bei einer falschen Antwort soll die richtige Antwort angezeigt werden.
Sie können den Code aus der vorherigen Aufgabe übernehmen
Versuchen Sie, eine Frage mit der Funktion prompt zu stellen. (ausblenden)
AUSPROBIEREN
WEITER ZU SCHLEIFEN

Page 16:

e-Learning: JavaScript

TUM Educational Media Lab

Meine Notizen

noch 35 Minuten

Sie haben 15 von 24 Seiten bearbeitet.

Schleifen

- Ein weiteres oft genutztes Konzept beim Programmieren sind sogenannte Schleifen
- Schleifen wiederholen einen Teil des Programmes solange, wie eine bestimmte Bedingung wahr ist
- Der Befehl dafür lautet *while*. Die allgemeine Form von Schleifen lautet

```
while (Bedingung) {
// Code, der ausgeführt werden soll, solange die Bedingung wahr ist.
}
```

- Sehen Sie sich das untere Beispiel an. Wir definieren in Zeile 1 zunächst die Variable x mit dem Wert 1. Unsere Bedingung der Schleife lautet "x < 4". Der Code zwischen der öffnenden geschweiften Klammer und der schließenden gescheiften Klammer wird also solange ausgeführt, bis die Variable x nicht mehr kleiner als 4 ist. Im Code zwischen den geschweiften Klammern (Zeile 3 und 4), zeigen wir erst den Inhalt der Variable x an und erhöhen dann die Variable x um 1.
- Überlegen Sie, was die Ausgabe des unteren Codes ist. Klicken Sie auf Ausprobieren, um zu sehen, ob Sie den Code richtig interpretiert haben.

```
1 var x = 1;
2 while (x < 4) {
3     alert(x);
4     x = x+1;
5 }
```

AUSPROBIEREN

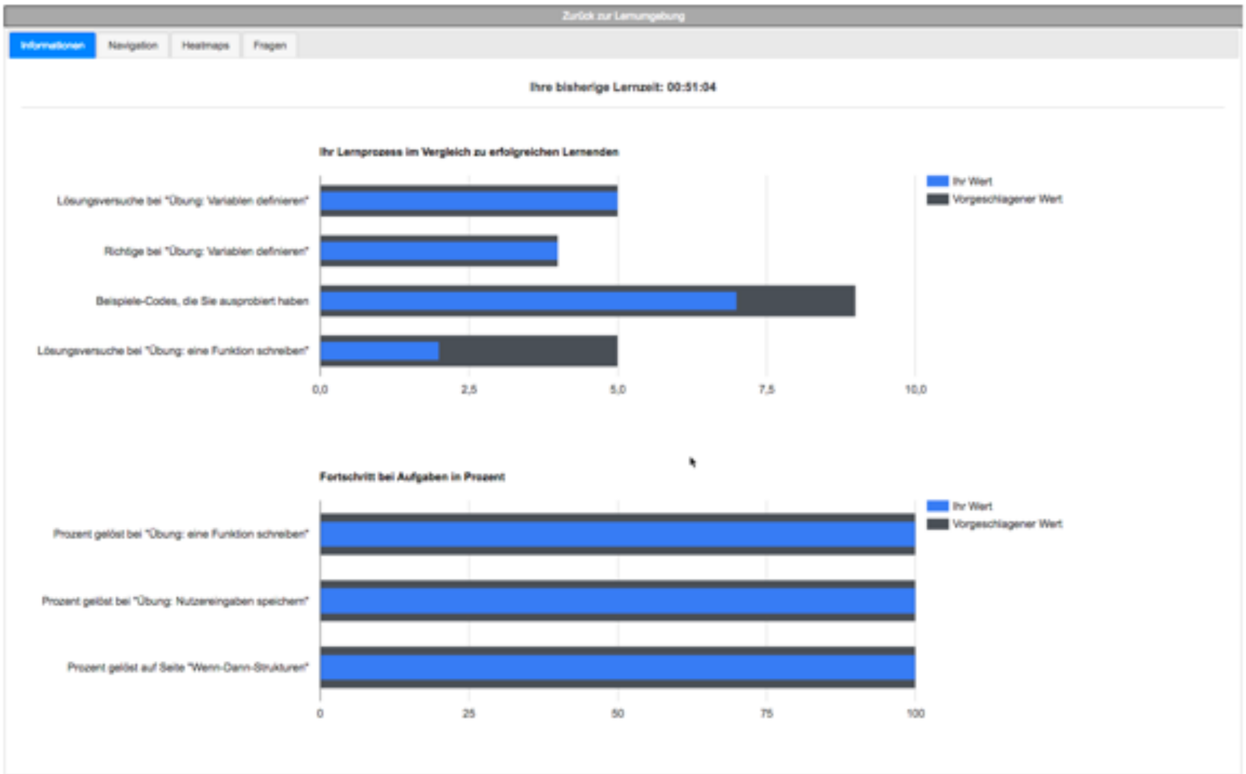
- Wichtig ist bei Schleifen noch, dass die Bedingung irgendwann nicht mehr erfüllt wird. Andernfalls wiederholt sich die Schleife unendlich oft - wir sprechen dann von einer Endlosschleife. Das würde zum Absturz unseres Programmes führen. Der folgende Code gibt ein Beispiel für eine Endlosschleife. Weil x bereits vor der Ausführung der Schleife 5 ist, kann die Bedingung nie erfüllt werden: x wird immer größer als 4 sein.

```
var x = 5;
while (x > 4) {
    alert(x);
    x = x+1;
}
```

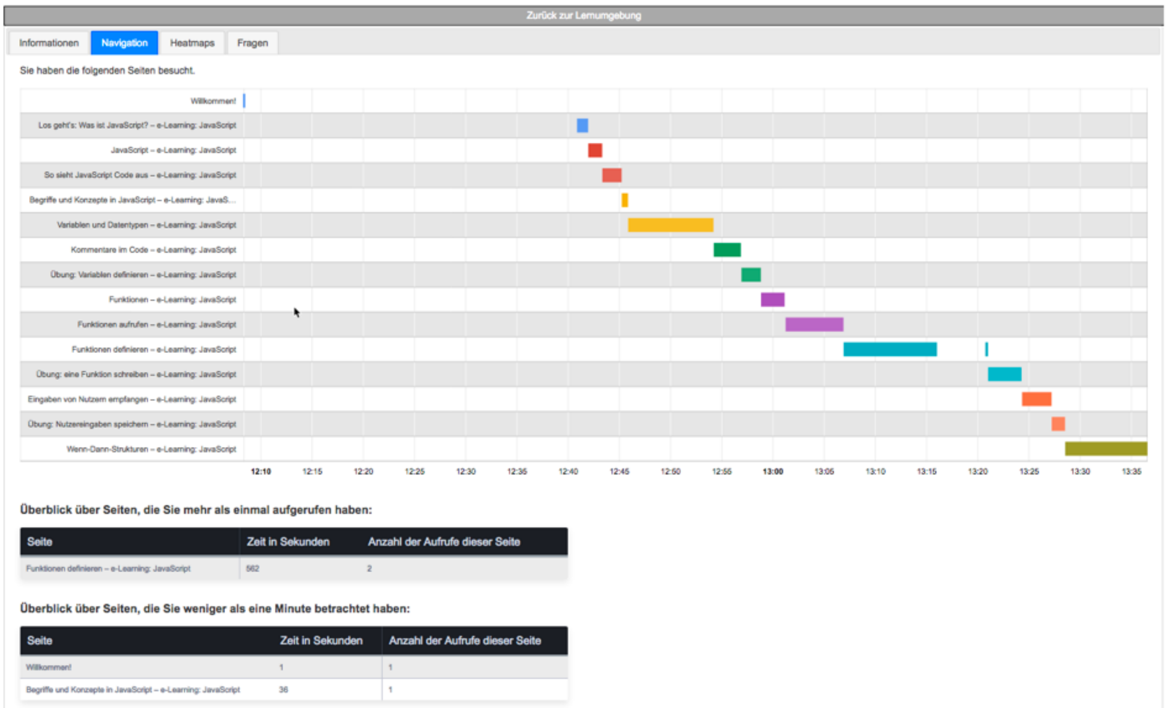
- Natürlich sind diese Programme nicht besonders nützlich. Eine praktischere Anwendung von Schleifen wäre in unserem Quiz-Beispiel denkbar. Dort könnten wir die Abfrage der Antwort solange wiederholen, bis ein Nutzer die richtige Antwort eingibt.
- Neben der While-Schleife gibt es auch noch die sogenannte for-Schleife. Die Besonderheit der for-Schleife ist, dass der Faktor, um den die Variable x in unserem Beispiel erhöht wird, direkt in der runden Klammer definiert wird.

Achtung: Dies ist die letzte inhaltliche Seite der Lernumgebung. Nachdem Sie auf den Weiter-Knopf gedrückt haben, können Sie nicht mehr auf die inhaltlichen Seiten zurückkehren, sondern gelangen zu den Fragebögen über Ihr Lernen.

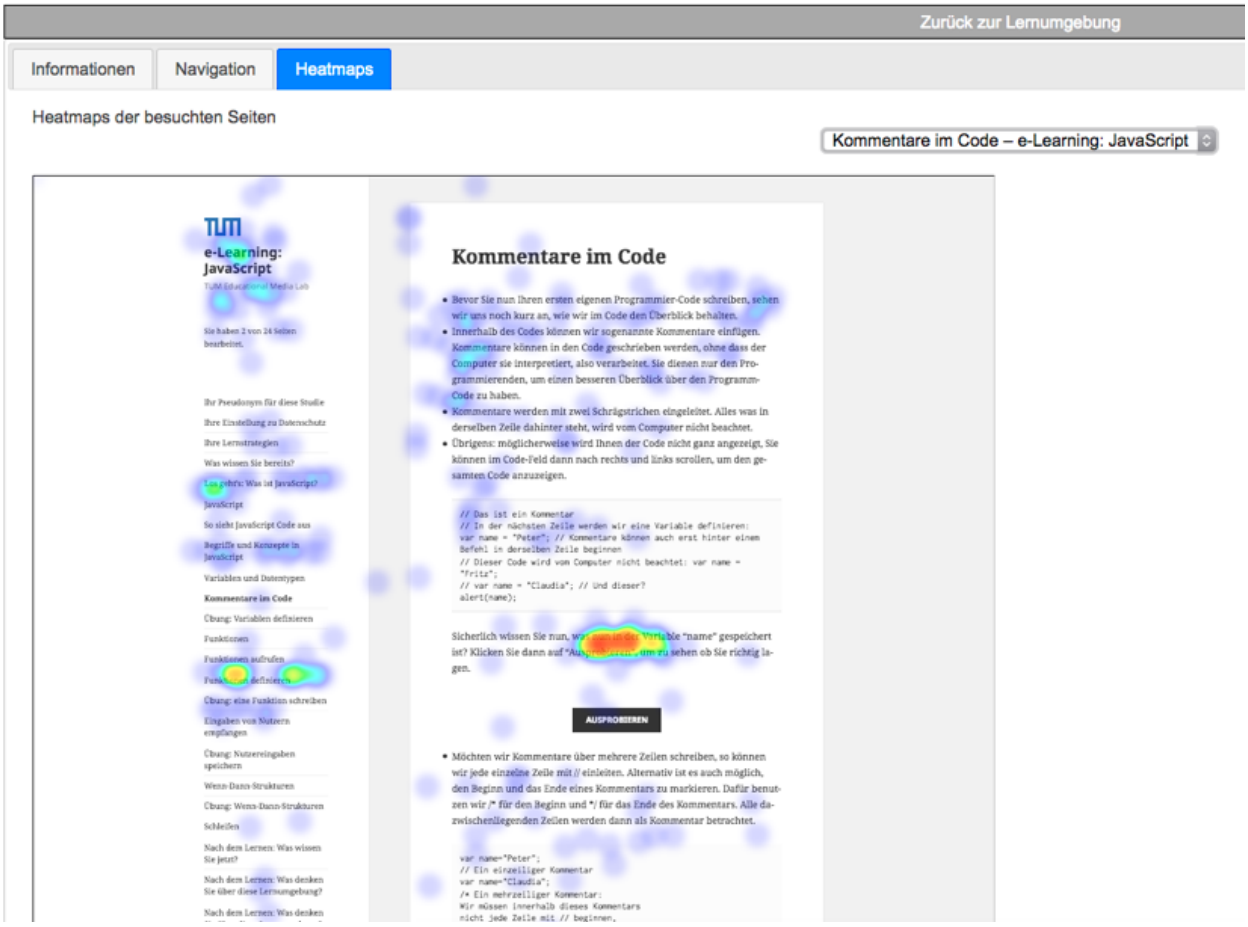
Dashboard Tab „Information“:



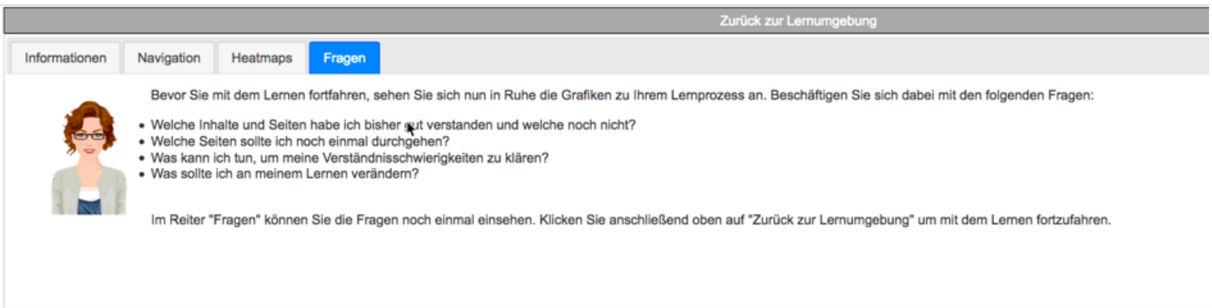
Dashboard Tab „Navigation“:



Dashboard Tab „Heatmaps“:



Dashboard Tab „Questions“:



Appendix B – Instructions

Appendix B1 – Instructions of Study 3

Herzlich Willkommen zur dieser Umfrage!

Vielen Dank für Ihr Interesse an dieser Studie teilzunehmen.
Um sicherzustellen, dass Sie die Voraussetzung für die Teilnahme und Vergütung in Höhe von 5 Euro erfüllen, lesen Sie sich die folgende Beschreibung der Studie bitte genau und vollständig durch. Wenn Sie Fragen zu diesem Projekt haben, wenden Sie sich gerne an Herrn Markus Hörmann.

Daten: Wir befragen Sie in dieser Studie zu verschiedenen Aspekten Ihrer Persönlichkeit und erfassen Daten zu Ihrem Antwortprozess. Zudem fragen wir Sie kurz nach Ihrem Vorwissen über einige gesellschaftlich relevante Themen. Zu Beginn der Befragung bitten wir Sie, Angaben zu Geschlecht, Alter und Beruf / Ausbildung / Studium zu machen.

Ablauf und Voraussetzungen: Die Beantwortung der Fragebögen wird **etwa 30 Minuten** in Anspruch nehmen. Zur Bearbeitung benötigen Sie einen Laptop oder Computer. **Eine Teilnahme mit mobilen Geräten (Tablet oder Handy) ist aus technischen Gründen leider nicht möglich. Bitte verlassen Sie diese Webseite bis zum Abschluss der Studie nicht, laden Sie die Seite nicht neu und nutzen Sie nicht die Vor- oder Zurück-Knöpfe Ihres Browsers - dies führt zum Verlust der bisher eingegebenen Antworten.** Vermeiden Sie insbesondere Ablenkungen durch Zugriffe auf Webseiten wie Facebook oder Google. Dies würde unsere Ergebnisse verfälschen. Bitte achten Sie darauf, dass Ihr Browser-Fenster den ganzen Bildschirm einnimmt (also maximiert / "groß geschalten" ist), da die Darstellung sonst nicht optimal ist.

Vergütung und Anonymisierung: Ihre Teilnahme an dieser Studie wird mit 5 Euro entschädigt. Bitte beachten Sie, dass eine Auszahlung der Vergütung nur bei vollständiger und sinnvoller Bearbeitung erfolgen kann. Um Missbrauch der Zahlungen zu verhindern, werden Ihre Daten vor Auszahlung auf **Vollständigkeit und sinnvolle Bearbeitung** geprüft. Damit wir die Auszahlung vornehmen können, fragen wir Sie am Ende der Erhebung nach Ihrer **Kontoverbindung**. Alternativ ist es möglich, den Betrag an den gemeinnützigen, entwicklungspolitischen Verein *Initiative Teilen im Cusanuswerk e.V.* zu spenden. Um die Anonymisierung Ihrer Daten zu gewährleisten, werden die Informationen zu Ihrer Bankverbindung nach erfolgter Zahlung und vor der Auswertung der Daten von uns vernichtet. Damit ist uns kein Rückschluss auf Ihre Person möglich.

Diese Studie wird durchgeführt von:
Markus Hörmann
Teaching and Learning with Digital Media
Technical University of Munich, TUM School of Education

Appendix B2 – Instructions of Study 4

Herzlich Willkommen!

Die Lernumgebung, in der Sie gleich arbeiten werden, öffnet sich in einem neuen Fenster. Bitte achten Sie darauf, dass das Fenster "groß geschalten" ist, da die Darstellung sonst nicht optimal ist.
Klicken Sie auf "Studie starten" um zu beginnen.

Vielen Dank, dass Sie an dieser Studie teilnehmen möchten!
Bitte lesen Sie die folgende Beschreibung der Studie genau und vollständig durch, um sicherzustellen, dass Sie die Voraussetzung für die Teilnahme und Vergütung in Höhe von 10 Euro erfüllen. Wenn Sie Fragen zu diesem Projekt haben, wenden Sie sich bitte an Herrn Markus Hörmann.

Wir, der Lehrstuhl für Lehren und Lernen mit digitalen Medien der Technischen Universität München, testen in dieser Studie eine Lernumgebung zum Thema "Einführung in die Programmierung von JavaScript". Der Test dient dazu, die Wirksamkeit und Qualität der Lehrmaterialien zu überprüfen.
Die Lernumgebung richtet sich gezielt an **Lernende ohne oder mit sehr geringem Vorwissen über das Programmieren**. Nehmen Sie bitte nicht teil, falls Sie bereits mit der Programmierung von JavaScript oder einer anderen Programmiersprache vertraut sind. Bitte nehmen Sie auch nicht teil, falls Sie in der Vergangenheit bereits mit dieser Lernumgebung gearbeitet haben.
Wir erheben in dieser Studie Daten zu Ihrem Lernprozess, zu Ihrem Wissen und befragen Sie zu Ihren Lernstrategien. Außerdem werden wir Sie zu Beginn bitten, einige Angaben zu Geschlecht, Alter und Ihrer Ausbildung / Ihrem Studium zu machen. Ihre Daten werden **vollständig pseudonymisiert behandelt**, d.h. wir können keine Rückschlüsse auf Ihre Person ziehen.

Die Bearbeitung der Lernumgebung wird **etwa 60 Minuten dauern**. Zudem wird das Ausfüllen von Fragebögen **etwa 15 bis 20 Minuten** Ihrer Zeit in Anspruch nehmen.
Am Ende der Sitzung können Sie auswählen, ob der Betrag per Überweisung an Sie ausbezahlt werden soll oder an den gemeinnützigen, entwicklungspolitischen Verein *Initiative Teilen im Cusanuswerk e.V.* gespendet werden soll.

Die Lernumgebung, in der Sie gleich arbeiten werden, öffnet sich in einem neuen Fenster. **Bitte achten Sie darauf, dass das Fenster den ganzen Bildschirm einnimmt (also maximiert / "groß geschalten" ist)**, da die Darstellung sonst nicht optimal ist.

Bitte bestätigen Sie noch einmal die folgenden Bedingungen zur Teilnahme durch Setzen der Haken und klicken Sie anschließend auf "Lernumgebung starten".

- [ ] Ich habe jetzt mindestens 75 Minuten Zeit, um ungestört und ohne Unterbrechung an dieser Studie teilzunehmen.
- [ ] Ich arbeite an einem Desktop-Computer oder Laptop, nicht an einem Tablet oder Handy.
- [ ] Ich habe kein oder sehr geringes Vorwissen zum Thema Programmieren.
- [ ] Ich habe noch nicht mit dieser Lernumgebung gearbeitet.
- [ ] Ich bin damit einverstanden, dass pseudonymisierte Daten zu meinem Lernprozess und Wissen und meinen Lernstrategien erfasst werden.
- [ ] Ich bin über 18 Jahre alt und besitze ein Bankkonto für die Überweisung der Vergütung in Höhe von 10 Euro.

Lernumgebung starten

Appendix C – Learning Tests

Appendix C1 – Learning Tests of Study 1 and 2

Declarative knowledge:

Wissen Sie bereits etwas über CSS?

- Die nächsten Fragen sollen Ihr Vorwissen über CSS erfassen.
- Hatten Sie bereits mit der Entwicklung von Webseiten in CSS zu tun?
  - Ja
  - Nein
- Falls ja: beschreiben Sie Ihre Vorerfahrung bitte kurz

  Ich habe keine Vorerfahrung

  - Wissen Sie wofür CSS steht? Tragen Sie Ihren Lösungsvorschlag in das folgende Feld ein:

    Ich weiß es nicht

  - Wie wird die externe Stylesheet-Datei center.css eingebettet?
    - <style type="text/css" src="center.css" />
    - <link href="center.css" type="stylesheet" />
    - <style src="center.css" rel="stylesheet" />
    - <link rel="stylesheet" href="center.css" />
    - Ich weiß es nicht

  - Welche der folgenden Eigenschaften verändern die Schrift?
    - font-type
    - font-color
    - font-align
    - font-style
    - Ich weiß es nicht

- Welche Aussagen sind korrekt?
  - ☐ ID-Selektoren können mehrere Elemente auswählen
  - ☐ Klassen-Selektoren können mehrere Elemente auswählen
  - ☐ Tag-Selektoren gelten für alle Elemente der Seite
  - ☐ ID-Selektoren können direkt in das öffnende Tag geschrieben werden
  - ☐ Ich weiß es nicht

- Welche Aussagen sind korrekt?
  - ☐ Eigenschaften müssen mit Klammern geöffnet und geschlossen werden
  - ☐ Werte können in verschiedenen Einheiten angegeben werden
  - ☐ Selektoren stehen zwischen geschweiften Klammern
  - ☐ Nach jedem Selektor steht ein Strichpunkt
  - ☐ Ich weiß es nicht

- Welche Aussagen sind richtig?
  - ☐ Eine Eigenschaft kann mehrerer Werte besitzen.
  - ☐ Klassen-Attribute dürfen nur einmal vergeben werden.
  - ☐ ID-Attribute werden mit einem # ausgewählt
  - ☐ Mit einem Element-Selektor können alle Elemente desselben Typs gleichzeitig verändert werden
  - ☐ Ich weiß es nicht

- Welche Aussagen sind richtig?
    - ☐ Das Attribut um in öffnenden Tags Stylesheets zu schreiben lautet "style"
    - ☐ Externe CSS-Dateien müssen mit <style> beginnen
    - ☐ Das Tag um innerhalb von HTML-Dateien CSS-Code zu schreiben lautet <style>
    - ☐ Die Dateiendung von externen CSS-Dateien heißt .style
    - ☐ Ich weiß es nicht

- Welche Aussagen sind richtig?
    - ☐ Ein Punkt steht für einen Klassenselektor, ein # für einen ID-Selektor
    - ☐ CSS kann auch als eigenständige Designsprache verwendet werden
    - ☐ Element-Selektoren gelten für alle Elemente der gesamten Webseite
    - ☐ Ein # steht für einen Klassenselektor-Selektor, ein Punkt für einen ID-Selektor
    - ☐ Ich weiß es nicht

- Welche Eigenschaften existieren in CSS **nicht**?
    - ☐ Border-color
    - ☐ Text-color
    - ☐ margin-color
    - ☐ pending
    - ☐ Ich weiß es nicht

- Ein Container mit den Eigenschaften width: 300px, padding: 50px und margin: 100px ist
    - 350px breit
    - 450px breit
    - 400px breit
    - 150px breit
    - Ich weiß es nicht

- Ein Container mit den Eigenschaften height: 150px, padding-bottom: 30px; padding: 10px und margin: 20px ist
    - 200px hoch
    - 180px hoch
    - 190px hoch
    - 220px hoch
    - Ich weiß es nicht

- Welche Aussagen sind richtig? Im Box-Modell
    - bezeichnet die Eigenschaft margin den Innenabstand und padding den Außenabstand
    - bezeichnet die Eigenschaft margin den Außenabstand und padding den Innenabstand
    - können margin und padding für eine spezifische Richtung definiert werden, border jedoch nicht.
    - gehört auch die Größe der Box zu den veränderbaren Eigenschaften.
    - Ich weiß es nicht

- Welche der folgenden Aussagen über Container sind richtig?
  - ☐ Container können ähnlich wie ID-Selektoren gestaltet werden
  - ☐ Container können eine Webseite in Bereiche aufteilen
  - ☐ Container können andere HTML-Elemente beinhalten
  - ☐ Container sind häufig H1 und P-Elemente
  - ☐ Ich weiß es nicht

- Welche Aussagen über CSS-Eigenschaften treffen zu?
  - ☐ text-align kann die Werte top oder bottom annehmen
  - ☐ flow bestimmt, in welche Richtung Elemente positioniert werden
  - ☐ font-style kann Schrift fett formatieren
  - ☐ font-weight kann Schrift fett formatieren
  - ☐ Ich weiß es nicht

Procedural knowledge:

- Stellen Sie sich vor, Sie haben folgenden HTML-Code vor sich:

```
<h1>Tiere</h1>
<div><img src="squirrel.png" id="squirrel" /></div>
<h1>Musikinstrumente</h1>
<p class="saiten">Gitarre</p>
<p class="tasten">Klavier</p>
<p class="tasten">Melodika</p>
<p class="saiten">Geige</p>
<p class="saiten">Cello</p>
<p class="schlag">Xylophon</p>
<h1>Sportarten</h1>

```

- Schreiben Sie nun den passenden CSS-Code. Ihre Aufgabe ist es ...
  - Alle Überschriften auf Ebene 1 (<h1>) auf die Schriftgröße(font-size) 12px einzustellen
  - Der Eigenschaft "background-color" des Elements "body" den Wert "cornflowerblue" zuzuweisen
  - Das Bild mit der ID "squirrel" in der Breite (width) zu 50px zu verändern
  - Alle Aufzählungspunkte mit der Klasse "saiten" in der Schriftfarbe Blau (blue) zu markieren
  - Den Innenabstand im Container auf 0.5cm zu setzen

Ich weiß es nicht

Appendix C2 – Learning Tests of Study 4

Was wissen Sie bereits?

Wenn Sie noch nichts oder noch nicht viel über die Programmierung mit JavaScript wissen, sind Sie hier genau richtig. Sehen Sie sich trotzdem die folgenden Fragen an und versuchen Sie diese zu beantworten. Vielleicht können Sie sich ja doch bereits die ein oder andere Antwort aus einem anderen Gebiet ableiten. Wenn Sie die Antwort nicht wissen, kreuzen Sie bitte "Ich weiß es nicht" an oder geben Sie "KA" für "Keine Ahnung" ein. Es können stets eine oder mehrere Antworten richtig sein.

Welche Vorteile bieten dynamische Webseiten gegenüber statischen Webseiten?

- [ ] Enthalten bewegliche Elemente wie z.B. Videos oder Animationen
- [ ] Können flexibel auf verschiedenen Geräten angezeigt werden
- [ ] Können auch nach dem Ladevorgang noch verändert werden
- [ ] Erlauben die Veränderung von Inhalt, Struktur und Aussehen
- [ ] Ich weiß es nicht

Welche der folgenden Aussagen zu JavaScript sind richtig?

- [ ] JavaScript Codes sind meist Bestandteile von Webseiten
- [ ] JavaScript wird häufig auch abgekürzt Java genannt
- [ ] JavaScript kann vom Benutzer deaktiviert werden.
- [ ] Mit JavaScript hat der Programmierende Zugriff auf alle Inhalte einer Webseite wie z.B. Texte, Bilder oder Tabellen
- [ ] Ich weiß es nicht

Benennen Sie die Datentypen der folgenden Variablen (schreiben Sie KA falls Sie es nicht wissen):

var x = new Image(); [ ]

---

var s = 5.65; [ ]

---

var t = "AB";

---

var b = "false";

---

var c = true;

---

var n = ["birne", "apfel", "mango"];

Wie heißt der Datentyp, der zwar ganze Zahlen, aber keine Kommazahlen repräsentieren kann?

Wie lautet der Befehl, mit dem Sie eine Variable definieren?

Nennen Sie zwei Punkte, die bei der Benennung von Variablen beachtet werden müssen.

Welche der folgenden Aussagen sind falsch?

- [ ] Bei JavaScript kommt es auf Groß- und Kleinschreibung an
- [ ] Kommentare werden in JavaScript mit \\ eingeleitet
- [ ] Die Aussage (10 > 9) liefert einen Wert vom Typ Integer zurück
- [ ] Kommentare stehen immer zu Beginn einer Zeile
- [ ] Ich weiß es nicht

Welche Aussagen zu Funktionen sind richtig?

- [ ] Funktionen können mehr als ein Argument empfangen
- [ ] Beim Funktionsaufruf wird der Name der Funktion benötigt
- [ ] Funktionen können weniger als ein Argument empfangen
- [ ] Argumente von Funktionen stehen zwischen geschweiften Klammern.
- [ ] Ich weiß es nicht

Welche weiteren Aussagen zu Funktionen sind richtig?

- [ ] Funktionen helfen dabei, bereits geschriebenen Code nochmal zu verwenden
- [ ] Alle Funktionen, die benutzt werden, müssen auch definiert werden
- [ ] Jede Funktion gibt einen Wert zurück
- [ ] Argumente von Funktionen werden mit Strichpunkten (;) voneinander getrennt
- [ ] Ich weiß es nicht

Welche Aussagen zu Wenn-Dann-Strukturen sind richtig?

- [ ] Bei Wenn-Dann-Strukturen steht die Bedingung in runden Klammern
- [ ] Wo eine If-Struktur steht muss nicht immer eine else-Struktur stehen
- [ ] Else-Strukturen benötigen keine Bedingung
- [ ] In Wenn-Dann-Strukturen werden mehrere Bedingungen mit Kommata getrennt
- [ ] Ich weiß es nicht

Welche Aussagen zu Schleifen sind richtig?

- [ ] Schleifen benötigen immer eine Bedingung
- [ ] Schleifen führen Code solange aus, wie eine Bedingung erfüllt ist
- [ ] Schleifen können Bedingungen enthalten, die unendlich oft ausgeführt werden
- [ ] Mehrere Bedingungen in Schleifen werden mit Kommata getrennt
- [ ] Ich weiß es nicht

Sehen Sie sich den folgenden JavaScript Code an:

```
var x=2;
while (x<=5){
x = x*x;
}
alert(x);
```

Welchen Wert hat die Variable x nach Abarbeitung des Codes? (geben Sie KA ein, wenn Sie es nicht wissen)

x hat den Wert ________

Beschreiben Sie nun kurz in eigenen Worten, was in den Schritten des oben gezeigten Code passiert.

Definieren Sie bitte eine Funktion in JavaScript, die den Mittelwert von zwei Werten berechnet. Sie soll folgende Eigenschaften haben:

- Die Funktion hat den Namen *mittelwert*
- Die Funktion bekommt als Argumente die Werte a und b übergeben
- Die Funktion gibt beim Aufruf den Mittelwert der Argumente zurück
- Die Funktion beachtet, dass sich der Mittelwert zweier Argumente wie folgt berechnet: m = (a+b)/2

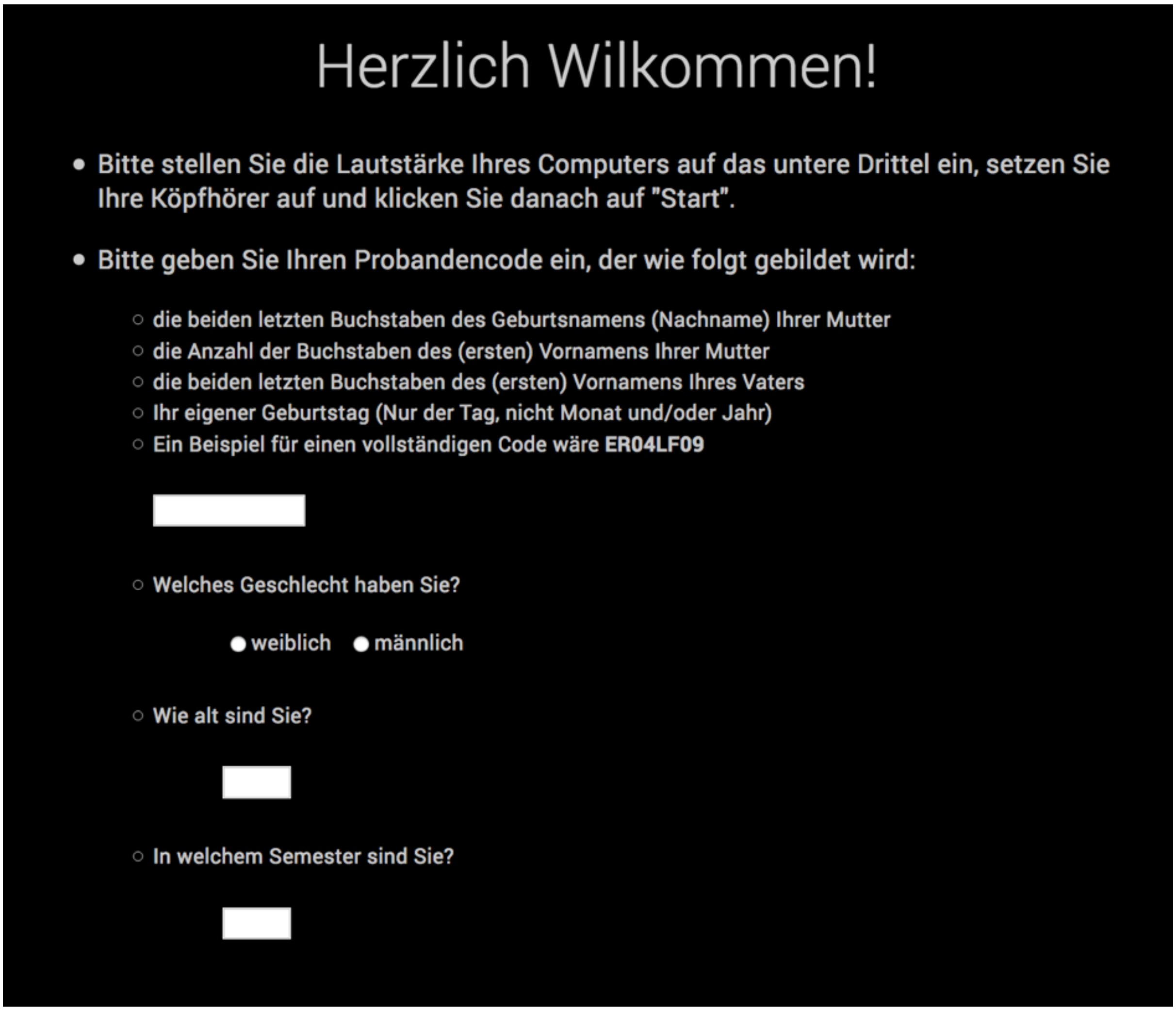
Benennen Sie alle Fehler im folgenden Code:

```
Var x=promt(Wie alt sind Sie?)
if [x=>18] {
alert(Volljährig!);
```

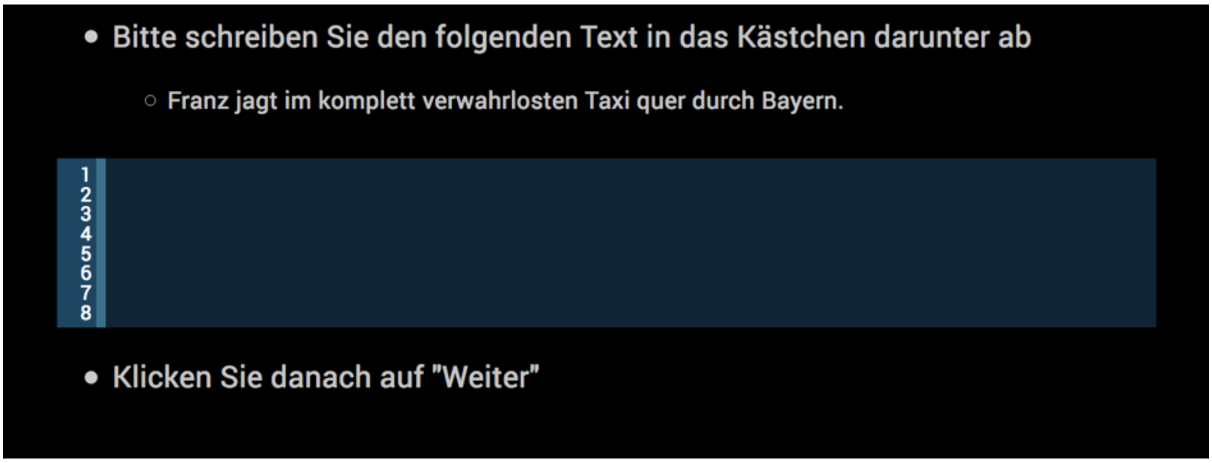
WEITER

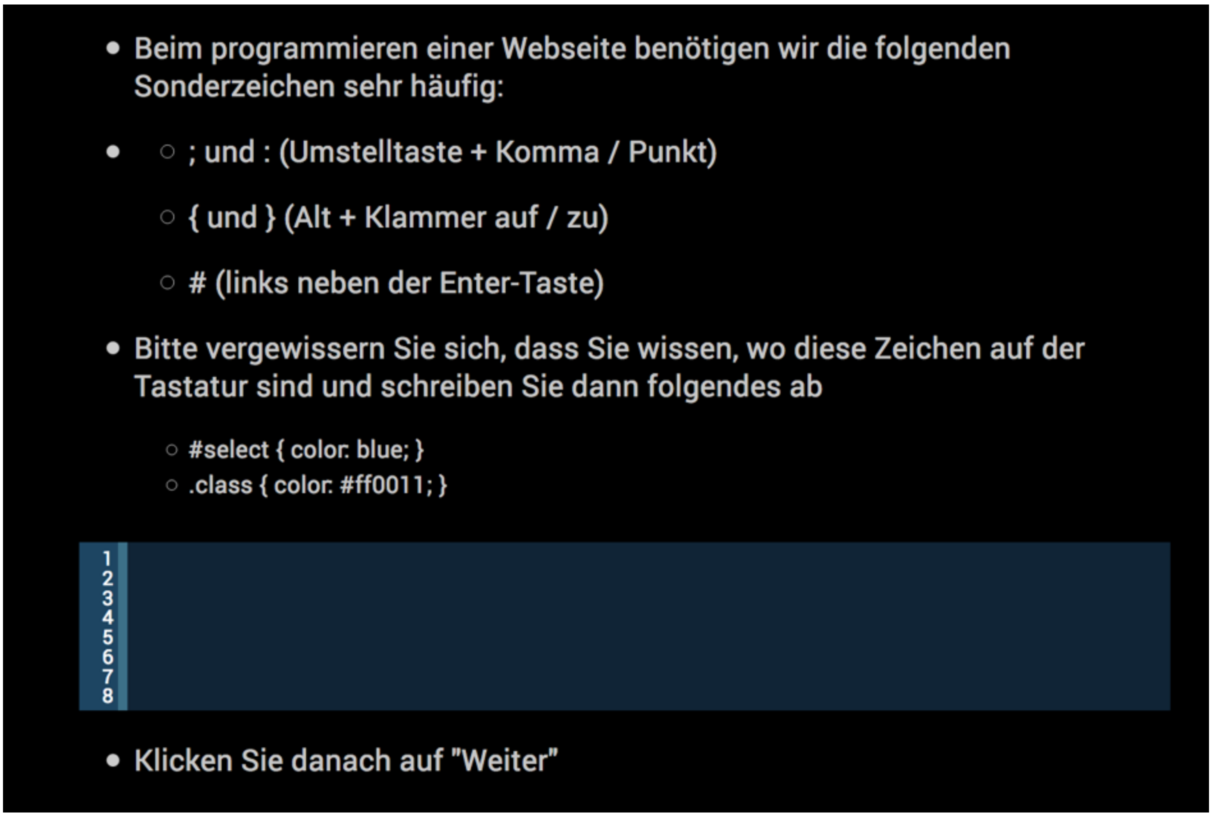
Appendix D – Instruments

Appendix D1 – Demographics in Study 1 and 2



Appendix D2 – Baseline of Typing Behavior in Study 1





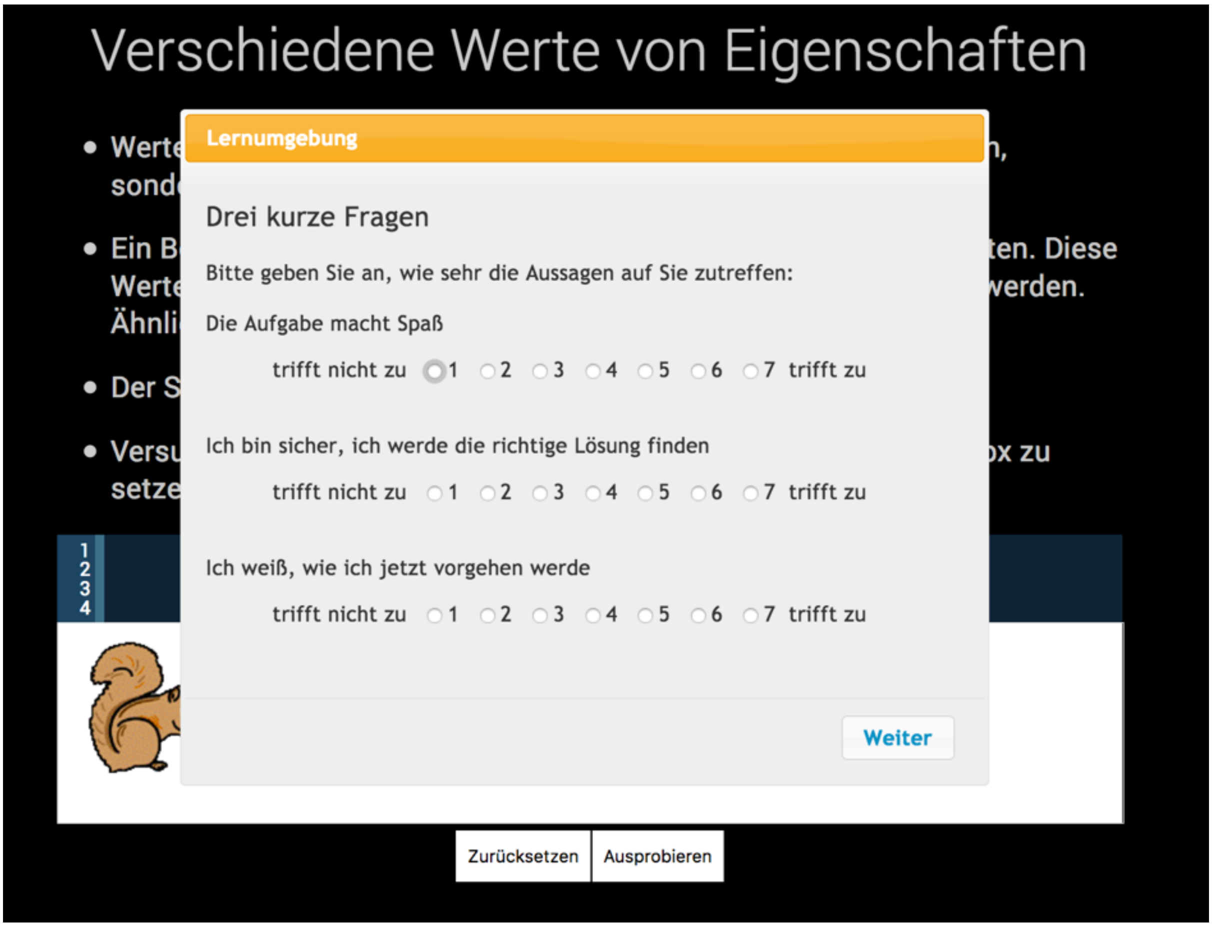
Appendix D3 – QCM for Initial Motivation in Study 1

Einige Fragen an Sie

- Bei dieser Lernumgebung wird es um das Lernen von HTML gehen, einer Sprache um Webseiten zu gestalten
- Dabei werden Sie auch selbst Code schreiben und an Aufgaben tüfteln
- Ausgehend davon, kreuzen Sie bitte an wie sehr die folgenden Aussagen auf sie zutreffen

	trifft nicht zu 1	2	3	4	5	6	trifft zu 7
Ich mag solche Rätsel und Knobeleien	●	●	●	●	●	●	●
Ich glaube, der Schwierigkeit dieser Aufgabe gewachsen zu sein.	●	●	●	●	●	●	●
Wahrscheinlich werde ich die Aufgabe nicht schaffen.	●	●	●	●	●	●	●
Bei der Aufgabe mag ich die Rolle des Wissenschaftlers, der Zusammenhänge entdeckt.	●	●	●	●	●	●	●
Ich fühle mich unter Druck, bei der Aufgabe gut abschneiden zumüssen.	●	●	●	●	●	●	●
Die Aufgabe ist eine richtige Herausforderung für mich.	●	●	●	●	●	●	●
Nach dem Lesen der Instruktion erscheint mir die Aufgabe sehr interessant.	●	●	●	●	●	●	●
Ich bin sehr gespannt darauf, wie gut ich hier abschneiden werde.	●	●	●	●	●	●	●
Ich fürchte mich ein wenig davor, dass ich mich hier blamieren könnte.	●	●	●	●	●	●	●
Ich bin fest entschlossen, mich bei dieser Aufgabe voll anzustrengen.	●	●	●	●	●	●	●
Bei Aufgaben wie dieser brauche ich keine Belohnung, sie machen mir auch so viel Spaß.	●	●	●	●	●	●	●
Es ist mir etwas peinlich, hier zu versagen.	●	●	●	●	●	●	●
Ich glaube, dass kann jeder schaffen.	●	●	●	●	●	●	●
Ich glaube, ich schaffe diese Aufgabe nicht.	●	●	●	●	●	●	●
Wenn ich die Aufgabe schaffe, werde ich schon ein wenig stolz auf mich sein.	●	●	●	●	●	●	●
Wenn ich an die Aufgabe denke, bin ich etwas beunruhigt.	●	●	●	●	●	●	●
Eine solche Aufgabe würde ich auch in meiner Freizeit bearbeiten.	●	●	●	●	●	●	●
Die konkreten Leistungsanforderungen hier lähmen mich.	●	●	●	●	●	●	●

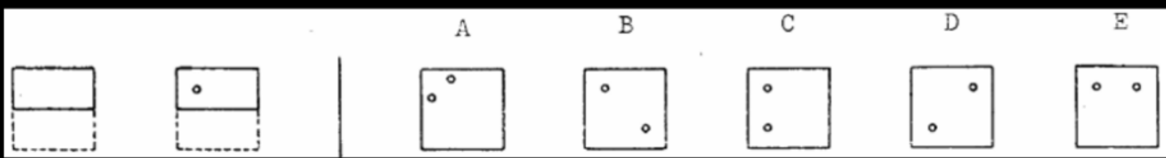
Appendix D4 – Adapted Short QCM for Current Motivation in Study 1



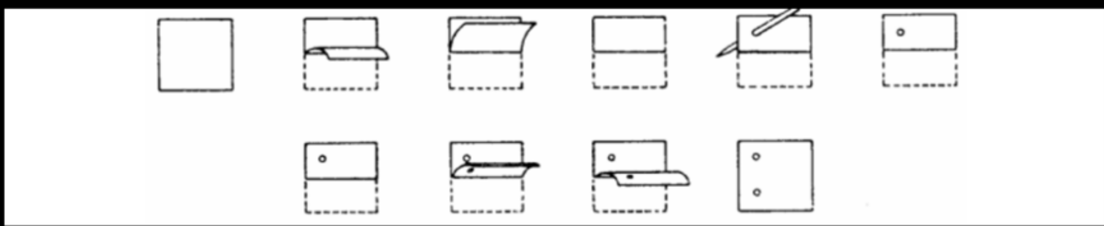
Appendix D5 – VZ-2 Paper Folding Test for Spatial Ability in Study 1

Ein kurzer Test zum räumlichen Vorstellungsvermögen

- Mit diesem Test möchten wir ihr räumliches Vorstellungsvermögen erfassen
- Dafür sollen Sie sich vorstellen, wie Papier gefaltet und wieder entfaltet wird
- In jeder Aufgabe sind ein paar Bilder links und rechts von einem Trennstrich abgebildet
- Diese Bilder stellen immer ein quadratisches Stück Papier dar, das gefaltet wird.
- Das letzte Bild auf der linken Seite hat immer einen Kreis abgebildet. Dort wurde das Papier durch alle gefalteten Schichten gelocht.
- Ein Bild auf der rechten Seite des Trennstrichs zeigt korrekt, wo sich die Löcher befinden, wenn das Papier wieder entfalten wird.
- Sehen Sie sich das folgende Beispiel an. Was könnte die Lösung sein?

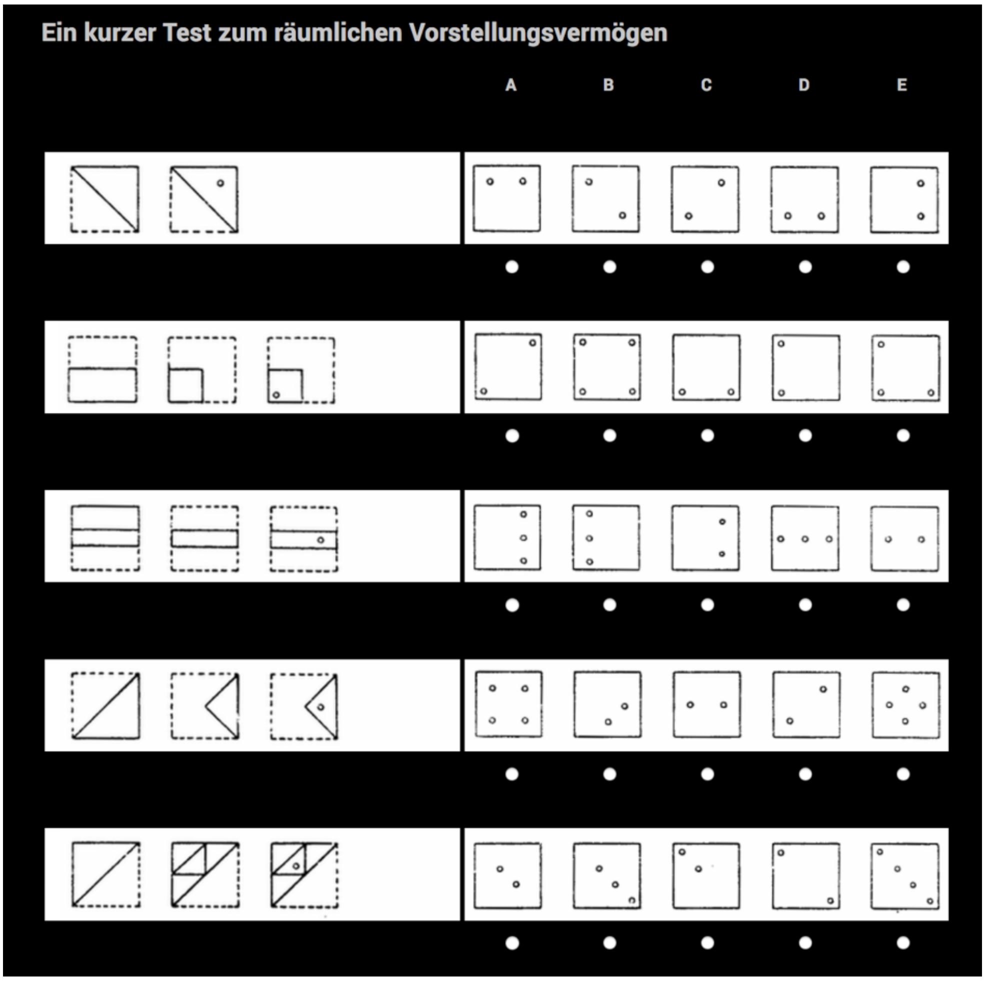


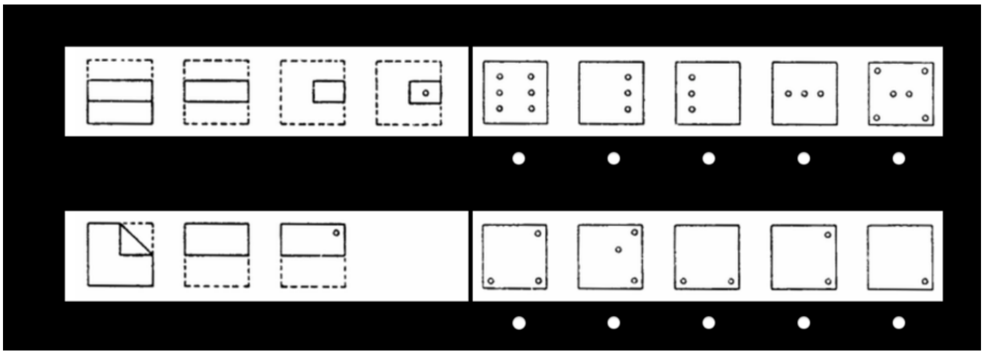
- Die korrekte Lösung ist C. Das Papier wurde wie folgt gefalten, gelocht und wieder aufgefalten.

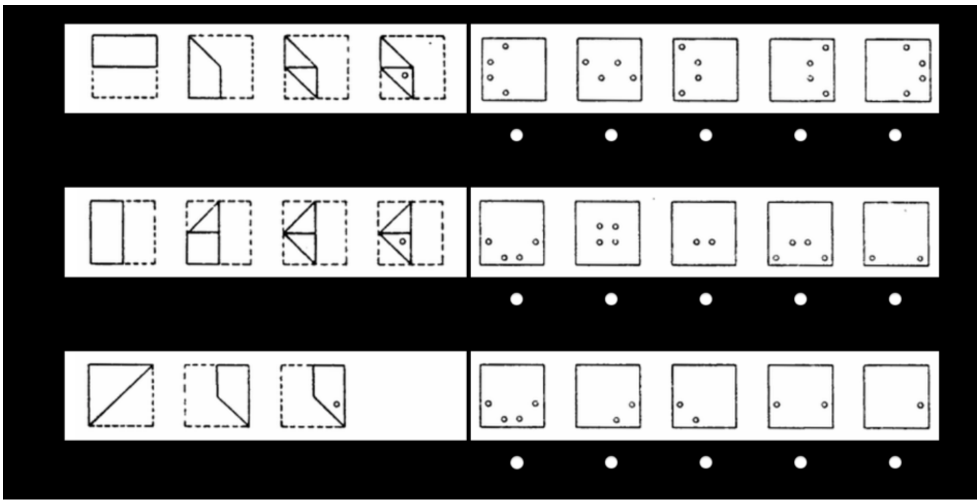


- Sie sollen nun in maximal 3 Minuten zehn dieser Aufgaben möglichst korrekt lösen
- Klicken Sie auf den "Start" Knopf, wenn Sie bereit sind, mit dem Test zu beginnen"

Start


Ein kurzer Test zum räumlichen Vorstellungsvermögen
A
B
C
D
E





Appendix D6 – PANAS for Affect in Study 2

Bevor es los geht: wie fühlen Sie sich?

- Im folgenden finden Sie eine Reihe von Wörtern, die unterschiedliche Gefühle und Empfindungen beschreiben. Lesen Sie jedes Wort und stellen Sie dann den entsprechenden Regler so ein, dass er auf Sie zutrifft. Es gibt hierbei keine richtigen oder falschen Antworten.
  Bitte geben Sie ehrlich an, wie Sie sich in diesem Moment fühlen.

aktiv	trifft überhaupt nicht zu	trifft voll und ganz zu
bekümmert	trifft überhaupt nicht zu	trifft voll und ganz zu
interessiert	trifft überhaupt nicht zu	trifft voll und ganz zu
freudig erregt	trifft überhaupt nicht zu	trifft voll und ganz zu
verärgert	trifft überhaupt nicht zu	trifft voll und ganz zu
stark	trifft überhaupt nicht zu	trifft voll und ganz zu
schuldig	trifft überhaupt nicht zu	trifft voll und ganz zu
erschrocken	trifft überhaupt nicht zu	trifft voll und ganz zu
feindselig	trifft überhaupt nicht zu	trifft voll und ganz zu
angeregt	trifft überhaupt nicht zu	trifft voll und ganz zu
stolz	trifft überhaupt nicht zu	trifft voll und ganz zu
gereizt	trifft überhaupt nicht zu	trifft voll und ganz zu
begeistert	trifft überhaupt nicht zu	trifft voll und ganz zu
beschämt	trifft überhaupt nicht zu	trifft voll und ganz zu
wach	trifft überhaupt nicht zu	trifft voll und ganz zu
nervös	trifft überhaupt nicht zu	trifft voll und ganz zu
entschlossen	trifft überhaupt nicht zu	trifft voll und ganz zu
aufmerksam	trifft überhaupt nicht zu	trifft voll und ganz zu
durcheinander	trifft überhaupt nicht zu	trifft voll und ganz zu
ängstlich	trifft überhaupt nicht zu	trifft voll und ganz zu

Appendix D6 – Demographics in Study 3

Ihr Alter:

Ihr Geschlecht

Männlich

Weiblich

Keine Angaben

Ihr Beruf / Studium / Ausbildung (ggf. welches Fach):

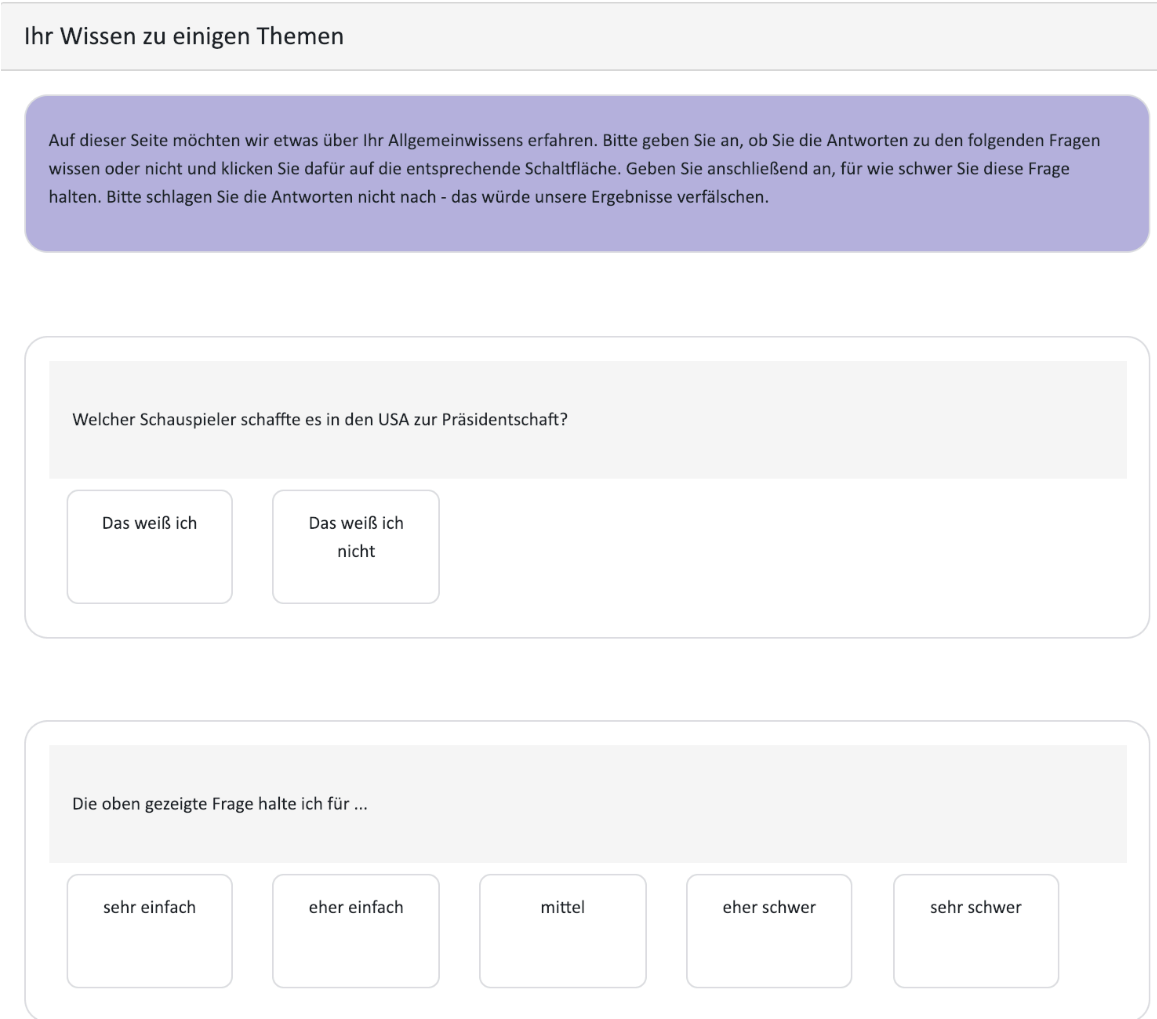
Teilnahmebedingungen:

- [ ] Ich habe jetzt mindestens 30 Minuten Zeit, um ungestört und ohne Unterbrechung an dieser Studie teilzunehmen.
- [ ] Ich arbeite an einem Desktop-Computer oder Laptop, nicht an einem Tablet oder Handy und mein Browser-Fenster ist maximiert.
- [ ] Ich habe noch nicht zuvor an dieser Studie teilgenommen.
- [ ] Ich bin damit einverstanden, dass anonymisierte Daten zu meinen Antworten erfasst werden.
- [ ] Ich bin über 18 Jahre alt und besitze ein Bankkonto für die Überweisung der Vergütung in Höhe von 5 Euro.
- [ ] Ich werde die Webseite dieser Studie bis zum Ende der Erhebung nicht verlassen.

Weiter

Appendix D7 – Design of Adapted BEFKI Judgements in Study 3

All items shown in "Appendix D8 – Items of the Adapted BEFKI Judgements in Study 3" were presented in the same structure and design as this example item.



Ihr Wissen zu einigen Themen

Auf dieser Seite möchten wir etwas über Ihr Allgemeinwissens erfahren. Bitte geben Sie an, ob Sie die Antworten zu den folgenden Fragen wissen oder nicht und klicken Sie dafür auf die entsprechende Schaltfläche. Geben Sie anschließend an, für wie schwer Sie diese Frage halten. Bitte schlagen Sie die Antworten nicht nach - das würde unsere Ergebnisse verfälschen.

Welcher Schauspieler schaffte es in den USA zur Präsidentschaft?

Das weiß ich

Das weiß ich nicht

Die oben gezeigte Frage halte ich für ...

sehr einfach

eher einfach

mittel

eher schwer

sehr schwer

Appendix D8 – Items of the Adapted BEFKI Judgements in Study 3

German (original)	English (translated)
Welcher Schauspieler schaffte es in den USA zur Präsidentschaft?	Which actor made it to the presidency in the USA?
Aus wie vielen Bundesländern besteht Deutschland?	How many federal states does Germany consist of?
Wie heißt der „Zeichentrick-Elefant“ mit den großen Ohren?	What is the name of the "cartoon elephant" with the big ears?
Welche Symptomatik ist typisch für Epilepsie	Which symptoms are typical for epilepsy?
Was war die Aufgabe der Inquisitionsgerichte des Mittelalters?	What was the task of the Inquisition courts of the Middle Ages?
Woraus besteht Bernstein?	What is amber made of?
Auf einem bekannten Gemälde von Dalí werden „zerfließende Uhren“ dargestellt. Welcher Stilrichtung ist dieses Gemälde zuzuordnen?	A well-known painting by Dalí depicts "melting clocks". What is the style of this painting?
Wozu dient Mitose?	What is Mitose for?
Was ist eine Petition?	What is a petition?
Was versteht man unter „Nihilismus“?	What is "nihilism"?
Welche Eigenschaft kennzeichnet eine Diode?	What is the characteristic of a diode?
Was ist das „Nibelungenlied“?	What is the "Nibelungenlied"?
Was sind Tantiemen?	What are royalties?
Familien- und Erbrecht sind Gegenstand welches Gesetzbuches?	Family law and inheritance law are the subject of which code?
Was passierte nach der „Völkerschlacht bei Leipzig“?	What happened after the "Battle of Leipzig"?
Wer erfand 1878 das Mikrofon?	Who invented the microphone in 1878?
Wann begann die Ausstrahlung des Farbfernsehens in der BRD?	When did the broadcast of colour television in Germany begin?
In welcher Stadt wurde Marilyn Monroe geboren?	In which city was Marilyn Monroe born?

Appendix D9 – Adapted BEFKI Answers in Study 3

Ihr Wissen zu einigen Themen

Nun möchten wir uns nochmal den Fragen von vorhin widmen. Im Folgenden werden Ihnen Fragen zu unterschiedlichen Themen gestellt. Ihre Aufgabe besteht darin, aus vier vorgegebenen Antwortmöglichkeiten die richtige herauszufinden. Für alle Fragen gilt, dass es jeweils nur eine einzige richtige Antwort gibt. Wählen Sie bei jeder Frage die zutreffende Antwort aus indem Sie auf die entsprechende Schaltfläche klicken. Falls Sie die richtige Antwort nicht wissen, dann raten Sie. Bitte lassen Sie keine Frage aus und schlagen Sie die Antworten nicht nach - das würde unsere Ergebnisse verfälschen.

Welcher Schauspieler schaffte es in den USA zur Präsidentschaft?

- Bill Clinton
- John Wayne
- Ronald Reagan
- Clint Eastwood

Aus wie vielen Bundesländern besteht Deutschland?

- 16
- 14
- 15
- 12

Wie heißt der „Zeichentrick-Elefant“ mit den großen Ohren?

- Balu
- Winnie Puuh
- Dumbo
- Bambi

Welche Symptomatik ist typisch für Epilepsie?

- Gedächtnisstörungen und Aufmerksamkeitsdefizite
- Krampfanfälle und Bewusstseinspausen
- lang anhaltende Schmerzen in den Gliedmaßen
- Übelkeit, Erbrechen und geistige Verwirrung

Was war die Aufgabe der Inquisitionsgerichte des Mittelalters?

- Entscheidungen in Rechtsfragen aller Art
- Durchsetzung des Volkswillens gegenüber der Feudalherrschaft
- Entscheidungen zu Fragen der Ethik und Moral
- Verurteilung von Ketzern und Hexen

Woraus besteht Bernstein?

- aus vulkanischem Magma
- aus fossilem Harz
- aus Silikaten
- aus Kristallen

Auf einem bekannten Gemälde von Dalí werden „zerfließende Uhren" dargestellt. Welcher Stilrichtung ist dieses Gemälde zuzuordnen?

- Naturalismus
- Impressionismus
- Surrealismus
- Romantik

Wozu dient die Mitose?

- Stoffwechselregulation
- Fortpflanzung
- Bildung von Keimzellen
- Zellvermehrung bei Wachstumsvorgängen

Was ist eine Petition?

- Einreichung einer Klage beim zuständigen Gericht
- Bitte oder Beschwerde an eine Behörde oder Volksvertretung
- Kandidatur für ein politisches oder soziales Amt
- Stellungnahme zu einem juristischen Sachverhalt

Was versteht man unter „Nihilismus“?

- Weltanschauung, die das Positive im Menschen betont
- Weltanschauung, die die Rolle der Moral betont
- Weltanschauung, die eine Sinnhaftigkeit der Welt bestreitet
- Weltanschauung, die den Erkenntnisgewinn als wichtigstes Prinzip ansieht

Welche Eigenschaft kennzeichnet eine Diode?

- Eine Diode lässt den elektrischen Strom nur in einer Richtung durch
- Eine Diode speichert elektrische Ladungen
- Eine Diode verstärkt elektrische Signale
- Eine Diode erzeugt ein Magnetfeld

Was ist das „Nibelungenlied“?

- bekanntes Gedicht von Friedrich Schiller
- aus der Antike überlieferte griechische Sage
- Nationalhymne der Schweiz
- mittelalterliches Heldenepos

Was sind Tantiemen?

- variable, umsatzabhängige Vergütungen
- Beiträge zur Sozialversicherung
- steuerliche Abgaben auf Lebensmittel
- Auszahlungen aus der Lebensversicherung

Familien- und Erbrecht sind Gegenstand des ...

- ... Bürgerlichen Gesetzbuches
- ... Sozialgesetzbuches
- ... Grundgesetzes
- ... Gemeinschaftsgesetzbuches

Nach der „Völkerschlacht bei Leipzig“ ...

- ... musste Kaiser Wilhelm II. abdanken
- ... rückten die alliierten Truppen nach Berlin vor
- ... musste sich Napoleon aus Deutschland zurückziehen
- ... wurde Karl der Große zum Kaiser gekrönt

Wer erfand 1878 das Mikrofon?

- Werner von Siemens
- Thomas Alva Edison
- Nikolaus August Otto
- David Edward Hughes

Wann begann die Ausstrahlung des Farbfernsehns in der BRD?

- 20. August 1968
- 25. August 1967
- 15. Juni 1965
- 21. August 1966

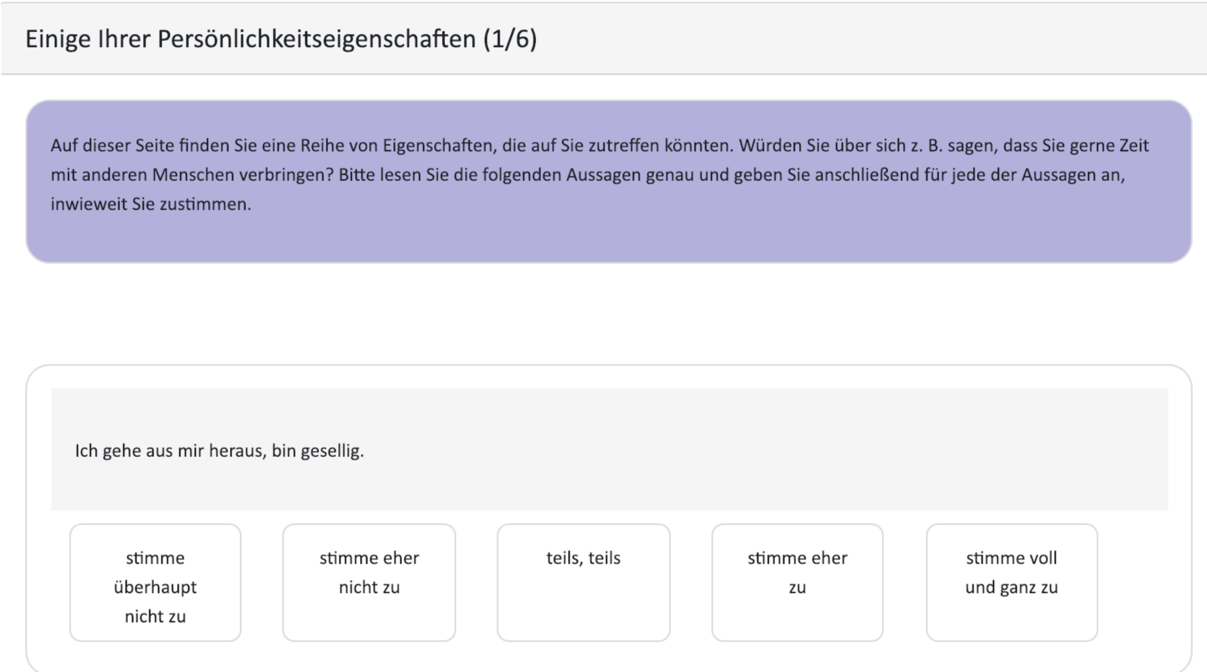
In welcher Stadt wurde Marilyn Monroe geboren?

- Seattle
- Houston
- New York
- Los Angeles

Weiter

Appendix D10 – Design of the Adapted BFI-2 with Confusion Induction in Study 3

*All items shown in "*Appendix D11 – Items of the Adapted BFI-2 with Confusion Induction in Study 3*" were presented in the same structure and design as this example item. 10 items were presented on one page.*


Einige Ihrer Persönlichkeitseigenschaften (1/6)

Auf dieser Seite finden Sie eine Reihe von Eigenschaften, die auf Sie zutreffen könnten. Würden Sie über sich z. B. sagen, dass Sie gerne Zeit mit anderen Menschen verbringen? Bitte lesen Sie die folgenden Aussagen genau und geben Sie anschließend für jede der Aussagen an, inwieweit Sie zustimmen.

Ich gehe aus mir heraus, bin gesellig.

stimme überhaupt nicht zu	stimme eher nicht zu	teils, teils	stimme eher zu	stimme voll und ganz zu

Appendix D11 – Items of the Adapted BFI-2 with Confusion Induction in Study 3

Page	German (original)	English (translated)	Manipulation
Page 1	Ich gehe aus mir heraus, bin gesellig.	I get out of myself, I'm sociable.	
	Ich bin einfühlsam, warmherzig.	I am sensitive, warm-hearted.	
	Ich bin eher unordentlich.	I am rather messy.	
	Ich bleibe auch in entspannten Situationen gelassen.	I stay calm even in relaxed situations.	Contradiction
	Ich bin nicht sonderlich kunstinteressiert.	I am not particularly interested in art.	
	Ich bin durchsetzungsfähig, energisch.	I am assertive, energetic.	
	Ich begegne anderen mit Respekt.	I treat others with respect.	
	Ich bin bequem, neige zu Faulheit.	I am comfortable, inclined to laziness.	
	Ich bleibe auch bei Rückschlägen zuversichtlich.	I remain confident even in the event of setbacks.	
	Ich bin vielseitig interessiert.	I am interested in many things.	
Page 2	Ich schäume selten vor Begeisterung über.	I seldom get too excited.	
	Ich neige dazu, andere zu kritisieren.	I tend to criticize others.	
	Ich bin stetig, beständig.	I am stable, steady.	
	Ich kann launisch sein, habe schwankende Stimmungen.	I can be moody, have fluctuating moods.	
	Ich bin erfinderisch, mir fallen raffinierte Lösungen ein.	I am inventive, I come up with sophisticated solutions.	
	Ich bin eher ruhig.	I am rather calm.	
	Ich habe mit anderen wenig Mitgefühl.	I have little sympathy with others.	
	Ich systematisch bin, halte mein Sachen in Ordnung.	I systematic am, keeping my things in order.	Grammar
	Ich reagiere leicht angespannt.	I react slightly tense.	
	Ich kann mich für Kunst, Musik und Literatur begeistern.	I can get enthusiastic about art, music and literature.	
Page 3	Ich neige dazu, die Führung zu übernehmen.	I tend to take the lead.	
	Ich habe oft Streit mit anderen.	I often quarrel with others.	
	Ich neige dazu, Aufgaben vor mir herzuschieben.	I tend to postpone tasks.	

	Ich bin selbstsicher, mit mir unzufrieden.	I am confident, dissatisfied with myself.	Contradiction
	Ich meide philosophische Diskussionen.	I avoid philosophical discussions.	
	Ich bin weniger aktiv und unternehmungslustig als andere.	I am less active and adventurous than others.	
	Ich bin nachsichtig, vergebe anderen leicht.	I am indulgent, forgiving others easily.	
	Ich bin manchmal ziemlich nachlässig.	I am sometimes quite careless.	
	Ich bin ausgeglichen, nicht leicht aus der Ruhe zu bringen.	I am balanced, not easily upset.	
	Ich bin nicht besonders einfallsreich.	I am not very imaginative.	
Page 4	Ich bin eher schüchtern.	I am rather shy.	
	Ich bin hilfsbereit und selbstlos.	I am helpful and selfless.	
	Ich mag es sauber und aufgeräumt.	I like it clean and tidy.	
	Ich mache mir oft Sorgen.	I am often worried.	
	Ich weiß Kunst und Schönheit zu schätzen.	I appreciate art and beauty.	
	Mir fällt es schwer, andere zu beeinflussen.	I find it hard to influence others.	
	Ich bin manchmal unhöflich und schroff.	I am sometimes rude and harsh.	
	Ich bin effizient, erledigt Dingen schnellen.	I am efficient, does things fast.	Grammar
	Ich fühle mich oft bedrückt, freudlos.	I often feel depressed, joyless.	
	Es macht mir Spaß, gründlich über komplexe Dinge nachzudenken und sie zu verstehen.	I enjoy thinking thoroughly about complex things and understanding them.	
Page 5	Ich bin voller Energie und Tatendrang.	I am full of energy and drive.	
	Ich bin anderen gegenüber misstrauisch.	I am suspicious of others.	
	Ich bin verlässlich, auf mich kann man zählen.	I am reliable, you can count on me.	
	Ich habe meine Gefühle unter Kontrolle, werde selten wütend.	I have my emotions under control, rarely get angry.	
	Ich bin nicht sonderlich fantasievoll.	I am not very imaginative.	
	Ich bin gesprächig.	I am talkative.	
	Andere sind mir eher gleichgültig, egal.	I don't care about anybody else.	

	Ich bin eher der ordentliche Typ, mache selten sauber.	I am rather the neat. I seldom clean up.	Contradiction
	Ich werde selten nervös und unsicher.	I rarely get nervous and insecure.	
	Ich finde Gedichte und Theaterstücke langweilig.	I find poems and plays boring.	
Page 6	In einer Gruppe überlasse ich lieber anderen die Entscheidung.	In a group I prefer to leave the decision to others.	
	Ich bin höflich und zuvorkommend.	I am polite and courteous.	
	Ich bleibe an einer Aufgabe dran, bis sie erledigt ist.	I stay on a task until it's done.	
	Ich bin oft deprimiert, niedergeschlagen.	I am often depressed, down.	
	Mich interessieren abstrakte Überlegungen wenig.	I am not interested in abstract considerations.	
	Ich bin begeisterungsfähig und kann andere leicht mitreißen.	I am enthusiastic and can easily carry others along with me.	
	Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen.	I easily trust others, believe in the good in people.	
	Manchmal verhalte mich verantwortunglos, leichtsinnig.	Sometimes act irresponsibly, reckless.	Grammar
	Ich reagiere schnell gereizt oder genervt.	I react quickly irritated or annoyed.	
	Ich bin originell, entwickle neue Ideen.	I am inventive, I develop new ideas.	

Appendix D12 – Demographics in Study 4

Bitte geben Sie nun Ihr Pseudonym und ihre Angaben zu Geschlecht, Alter und Ausbildung ein.

Ihr Pseudonym wird wie folgt gebildet:

- die beiden letzten Buchstaben des Geburtsnamens (Nachname) Ihrer Mutter
- die Anzahl der Buchstaben des (ersten) Vornamens Ihrer Mutter
- die beiden letzten Buchstaben des (ersten) Vornamens Ihres Vaters
- Ihr eigener Geburtstag (Nur der Tag, nicht Monat und/oder Jahr)

Ein Beispiel für einen vollständigen Code wäre ER04LF09. Geben Sie Zahlen bitte mit zwei Ziffern an, also z.B. 05 für 5.

Ihr Geschlecht

◯ männlich ◯ weiblich

Ihr Alter

Welchen Beruf oder welche Ausbildung/welches Studium üben Sie im Moment aus?

Haben Sie bereits Erfahrung mit Programmierung? Falls ja, welche?

Bitte geben Sie zur Sicherung der Datenqualität an, ob Sie nur einen Blick auf diese Studie werfen möchten oder an der Studie teilnehmen und sie vollständig bearbeiten werden.

◯ Ich möchte mir diese Studie nur ansehen.

◯ Ich möchte an dieser Studie teilnehmen und sie vollständig bearbeiten.

WEITER

Appendix D13 – Need for Privacy in Study 4

Beim Lernen mit digitalen Medien werden immer auch Daten zwischen Ihnen und dem Anbieter übermittelt. Daher möchten wir gerne von Ihnen wissen, was Sie über die folgenden Aussagen denken.

Bitte geben Sie für jede der folgenden Aussagen an, wie sehr diese auf Sie zutrifft:

Um meine Daten zu schützen, versuche ich keine Geräte zu nutzen, die Daten über mich sammeln (z.B. Fitness-Tracker).

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

Ich habe das Gefühl, mir fehlt das technische Verständnis, um für meinen Datenschutz zu sorgen.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

Ich bin besorgt darüber, dass elektronisch gespeicherte Daten über mich von anderen Personen gelesen werden könnten.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

Solange ein Service komfortabel ist, ist Datenschutz für mich nicht so wichtig.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

Ich installiere eine App nicht, wenn zu viele persönliche Daten verlangt werden.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

Bei Apps prüfe ich stets, welche Datenschutz-Einstellungen es gibt.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

Dass ich Spuren im Internet hinterlasse, beunruhigt mich.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Ich versuche mich vor Datenmissbrauch zu schützen (z.B. durch verschlüsselte Messenger, Firewalls, Löschen von Cookies, Nutzen eines Proxyservers).

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Ich versuche es zu vermeiden, meinen echten Namen im Internet anzugeben.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Das Thema Datenschutz ist mir im Allgemeinen sehr wichtig.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

WEITER

Appendix D14 – Adapted LIST for Metacognitive Strategies in Study 4

Im Folgenden möchten wir gerne mehr über Ihr gegenwärtiges Lernverhalten erfahren. Bitte geben Sie für jede der im folgenden genannten Aktivitäten die Häufigkeit an, mit der Sie diese üblicherweise ausführen, wenn Sie Lernen oder sich auf einen Test / eine Prüfung vorbereiten.

Wenn ich während des Lesens eines Textes nicht alles verstehe, versuche ich, die Lücken festzuhalten und den Text daraufhin noch einmal durchzugehen.

◯ nie ◯ selten ◯ gelegentlich ◯ oft ◯ immer

Ich kann nach einer Prüfung gut einschätzen, wie ich abschneiden werde.

◯ nie ◯ selten ◯ gelegentlich ◯ oft ◯ immer

Wenn ich einen schwierigen Text vorliegen habe, passe ich meine Lerntechnik den höheren Anforderungen an (z.B. durch langsameres Lesen).

◯ nie ◯ selten ◯ gelegentlich ◯ oft ◯ immer

Um mein eigenes Verständnis zu prüfen, erkläre ich bestimmte Teile des Lernstoffs jemand anderem.

◯ nie ◯ selten ◯ gelegentlich ◯ oft ◯ immer

Vor dem Lernen eines Stoffgebiets überlege ich mir, wie ich am effektivsten vorgehen kann.

◯ nie ◯ selten ◯ gelegentlich ◯ oft ◯ immer

Ich versuche, mir vorher genau zu überlegen, welche Teile eines bestimmten Themengebiets ich lernen muss und welche nicht.

◯ nie ◯ selten ◯ gelegentlich ◯ oft ◯ immer

Um Wissenslücken festzustellen, rekapituliere ich die wichtigsten Inhalte, ohne meine Unterlagen zu Hilfe zu nehmen.

◯ nie ◯ selten ◯ gelegentlich ◯ oft ◯ immer

Ich bearbeite zusätzliche Aufgaben, um festzustellen, ob ich den Stoff wirklich verstanden habe.

◯ nie ◯ selten ◯ gelegentlich ◯ oft ◯ immer

Wenn mir eine bestimmte Textstelle verworren und unklar erscheint, gehe ich sie noch einmal langsam durch.

◯ nie ◯ selten ◯ gelegentlich ◯ oft ◯ immer

Ich stelle mir Fragen zum Stoff, um sicherzugehen, dass ich auch alles verstanden habe.

◯ nie ◯ selten ◯ gelegentlich ◯ oft ◯ immer

Ich überlege mir vor dem Lernen, in welcher Reihenfolge ich den Stoff durcharbeite.

◯ nie ◯ selten ◯ gelegentlich ◯ oft ◯ immer

Ich lege im Vorhinein fest, wie weit ich mit der Durcharbeitung des Stoffs kommen möchte.

◯ nie ◯ selten ◯ gelegentlich ◯ oft ◯ immer

WEITER

Appendix D15 – Evaluation of the Dashboard in Study 4

Um diese Lernumgebung zu verbessern sind wir auf Ihre Erfahrung beim Lernen angewiesen.

Wir interessieren uns dafür, ob und wie Ihnen die Informationen zu Ihrem Lernprozess geholfen haben. Damit meinen wir nicht die Hinweise von Anna zu Fehlern in Ihren Aufgabenlösungen, sondern die allgemeinen Informationen (Grafiken zu Lernzeit, Heatmaps,etc.) die Ihnen angezeigt wurden. Bitte geben Sie für die folgenden Aussagen an, wie sehr Sie diesen zustimmen.

Die Informationen über meinen Lernprozess ...

... haben mich dazu angeregt, über mein Lernen nachzudenken.

○ stimmt überhaupt nicht ○ wenig ○ mittelmäßig ○ ziemlich ○ sehr

... empfand ich als nervend.

○ stimmt überhaupt nicht ○ wenig ○ mittelmäßig ○ ziemlich ○ sehr

... konnte ich nutzen, um meine Lernaktivitäten entsprechend anzupassen.

○ stimmt überhaupt nicht ○ wenig ○ mittelmäßig ○ ziemlich ○ sehr

... haben mich vom Lernen abgelenkt.

○ stimmt überhaupt nicht ○ wenig ○ mittelmäßig ○ ziemlich ○ sehr

... haben nichts an meinem Lernen verändert.

○ stimmt überhaupt nicht ○ wenig ○ mittelmäßig ○ ziemlich ○ sehr

... waren weder störend noch hilfreich.

○ stimmt überhaupt nicht ○ wenig ○ mittelmäßig ○ ziemlich ○ sehr

… waren interessant.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

… haben letztendlich zu einem besseren Verständnis des Inhalts beigetragen.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Bitte beschreiben Sie noch kurz, welche Teile der Informationen (Allgemeine Informationen zum Lernprozess, Grafik zu Ihrem Navigationsverhalten, Heatmaps) zu Ihrem Lernprozess Sie als hilfreich / nicht hilfreich empfanden? Weshalb? Was würden Sie verändern? Was würden Sie beibehalten?

WEITER

Appendix D16 – Evaluation of the Learning Environment in Study 4

Um diese Lernumgebung zu verbessern sind wir auf Ihre Erfahrung beim Lernen angewiesen.

Wir interessieren uns dafür, wie Sie den Umgang mit der Lernumgebung empfanden. Bitte geben Sie an, wie sehr die folgenden Aussagen auf Sie zutreffen.

Ich würde so eine Lernumgebung gerne häufiger benutzen.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Die Lernumgebung war einfach zu bedienen.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Ich bräuchte technische Hilfe um diese Lernumgebung sinnvoll nutzen zu können.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Ich fand die Nutzung der Lernumgebung mühsam.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Die verschiedenen Funktionen des Systems haben sich sinnvoll ergänzt.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Ich denke, die meisten Menschen können den Umgang mit diesem System schnell lernen.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Die Lernumgebung war unnötig komplex.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Ich habe mich sicher im Umgang mit der Lernumgebung gefühlt.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Ich musste viel Neues lernen, bevor ich das System nutzen konnte.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Bitte beschreiben Sie zudem noch kurz, wie Sie die Rückmeldungen von Anna empfanden. Was würden Sie verändern? Was würden Sie beibehalten?

WEITER

Appendix D17 – Adapted Cognitive Load Scale in Study 4

Bitte bewerten Sie die folgenden Aussagen ebenfalls in Bezug auf die Informationen, die Ihnen zu Ihrem Lernprozess angezeigt wurden. Damit meinen wir nicht die Hinweise von Anna zu Fehlern in Ihren Aufgabenlösungen, sondern die allgemeinen Informationen (Grafiken zu Lernzeit, Heatmaps,etc.) die Ihnen angezeigt wurden.

Es war schwer, die zentralen Informationen miteinander in Verbindung zu bringen.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Ich habe mich angestrengt, nicht nur einzelne Informationen anzusehen, sondern auch den Gesamtzusammenhang zu verstehen.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Es ging mir beim Durchsehen der Informationen darum, alles richtig zu verstehen.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Ich musste viele Informationen gleichzeitig im Kopf behalten.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Beim Durchsehen der Informationen war ich mental angestrengt.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Das Durchsehen der Informationen hat mir Spaß gemacht.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Die Informationen zu nutzen war eine sehr komplexe Aufgabe.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

Die Darstellung der Informationen ist ungünstig, um mein Lernen nachzuvollziehen.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

---

WEITER

Appendix D18 – Evaluation of the Pedagogical Agent in Study 4

Bitte bewerten Sie die Rückmeldungen, die Anna Ihnen auf Ihre Aufgabenlösungen gegeben hat.

Anna hat meine Fehler richtig erkannt.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

Anna hat mich dazu gebracht, über das Gelernte zu reflektieren.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

Anna hat mir geholfen, mich vertieft in die Inhalte hinein zu denken.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

Annas Rückmeldungen waren hilfreich.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

Anna empfand ich als angenehm.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

Anna hat mir geholfen, meine Kenntnisse über den Inhalt zu verbessern.

◯ stimmt überhaupt nicht ◯ wenig ◯ mittelmäßig ◯ ziemlich ◯ sehr

WEITER