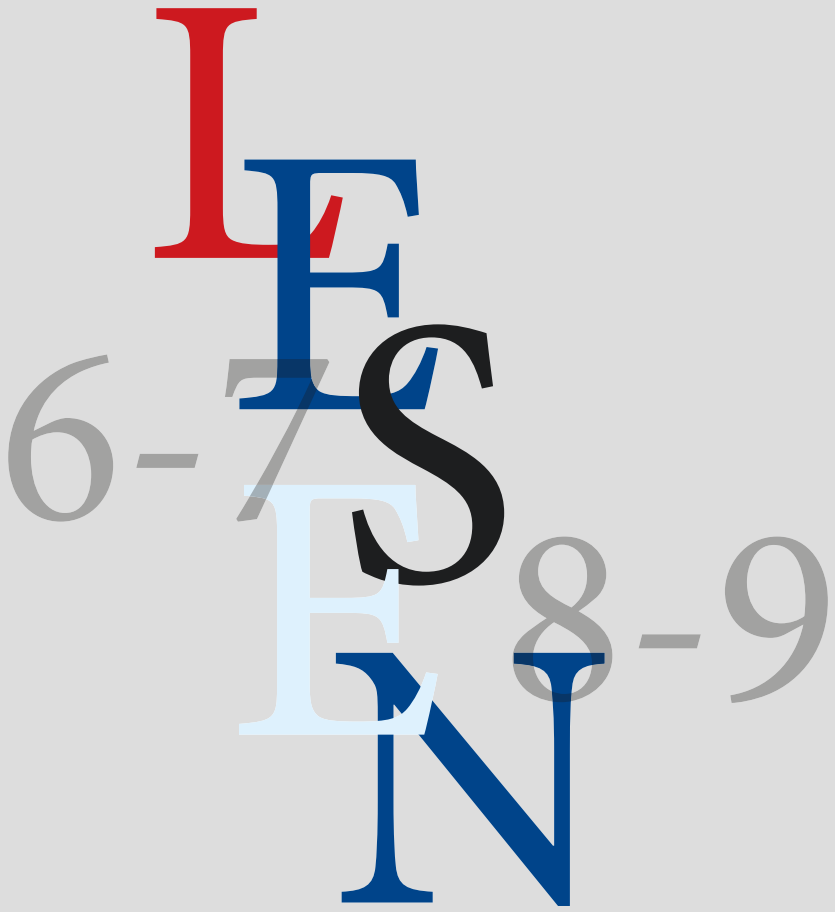


Kerstin Bäuerlein

Leseverständnisdiagnostik in der Sekundarstufe

Würzburg University Press

Kerstin Bäuerlein

Leseverständnisdiagnostik in der Sekundarstufe

Kerstin Bäuerlein

Leseverständnisdiagnostik in der Sekundarstufe

Theoretische Grundlagen sowie Konstruktion und empirische Erprobung der Lesetests LESEN 6-7 und LESEN 8-9

Würzburg University Press

Dissertation, Julius-Maximilians-Universität Würzburg
Philosophische Fakultät II, 2014
Gutachter: Prof. Dr. Wolfgang Schneider, PD Dr. Wolfgang Lenhard

Impressum

Julius-Maximilians-Universität Würzburg
Würzburg University Press
Universitätsbibliothek Würzburg
Am Hubland
D-97074 Würzburg
www.wup.uni-wuerzburg.de

Print on Demand

ISBN 978-3-95826-008-5 (print)
ISBN 978-3-95826-009-2 (online)
URN urn:nbn:de:bvb:20-opus-95329

Vorwort

Die vorliegende Arbeit beschäftigt sich mit der Frage, wie sich die Lesekompetenz in der Sekundarstufe entwickelt und welche Möglichkeiten verfügbar sind, sie auch zuverlässig zu erfassen. Seit der ersten Veröffentlichung der PISA-Ergebnisse im Jahr 2000 wissen wir, dass das Leseverständnis auch im Jugendalter vielfach noch nicht voll entwickelt ist. Während zur Diagnose der Lesekompetenz in der Grundschule mittlerweile zahlreiche Testverfahren entwickelt wurden, ist die Zahl der für die Sekundarstufe konzipierten Lesetests immer noch sehr überschaubar.

Diese Lücke soll künftig über zwei standardisierte Leseverfahren (LESEN 6-7 und LESEN 8-9) geschlossen werden, an deren Konstruktion und empirischer Erprobung Frau Bäuerlein maßgeblichen Anteil hatte. Wie aus der Darstellung der Entwicklungsarbeit im vorliegenden Buch klar hervorgeht, wurde das Projekt von Frau Bäuerlein vorbildlich konzipiert. Der theoretische Teil der Arbeit verdeutlicht, dass sich die Autorin tief in die verfügbare Literatur zur Entwicklung der Lesekompetenz eingearbeitet hat, weiterhin aber auch die Situation hinsichtlich der Diagnose von Lesekompetenz sehr gut überblickt. Diese Kenntnisse fließen dann konsequent in die Vorbereitung der eigenen Untersuchung ein. Es imponiert zum einen die Präzision der Planung, zum anderen aber auch die konsequente Umsetzung der Planungsschritte. So hat die Autorin einen beträchtlichen Aufwand betrieben, um die Testaufgaben für die beiden genannten Leseverfahren auszuwählen. Es wurden dazu nicht nur sorgfältig konzipierte Voruntersuchungen durchgeführt, sondern auch Lehrkräfte und Experten im Bereich der Leseforschung kontinuierlich in die Vorbereitung der Testversionen eingebunden. Damit sollte sichergestellt werden, dass die letztendlich ausgewählten Testaufgaben nicht nur den üblichen formalen Kriterien genügten, sondern auch inhaltlich voll akzeptabel waren. In theoretischer und methodischer Hinsicht genügt die Arbeit also höchsten Anforderungen.

Der Verfasserin ist es mit dem vorliegenden Werk gelungen, ein ambitioniertes Projekt zu einem erfolgreichen Abschluss zu bringen. Es werden Testverfahren zur Erfassung der Lesekompetenz in der Sekundarstufe vorgestellt, die hohen Ansprüchen genügen. Die wissenschaftliche Bedeutung der Arbeit ist insgesamt sehr hoch einzuschätzen, so dass ich dem Werk eine weite Verbreitung wünsche. Es bietet insbesondere für Leser aus dem Bereich der empirischen Bildungsforschung, der pädagogischen Psychologie und der Pädagogik eine Menge interessanter und nützlicher Informationen.

Würzburg, im November 2014 Prof. Dr. Wolfgang Schneider

Inhaltsverzeichnis

Einleitung

Die im Jahr 2000 erstmals durchgeführte internationale Schulleistungsvergleichsstudie PISA[1] führte in Deutschland zum sogenannten „PISA-Schock“. Die 15-jährigen deutschen Schüler[2] wiesen im Hinblick auf die Lesekompetenz zum Teil gravierende Defizite auf und schnitten im internationalen Vergleich unerwartet schlecht ab (Artelt, Stanat, Schneider & Schiefele, 2001). Das Gewicht dieser Befunde wird deutlich, wenn man sich vor Augen führt, dass Lesekompetenz trotz – oder gerade wegen – der neuen Medien heute bedeutsamer ist denn je. Sie ist eine Voraussetzung für die Aneignung weiterer Kompetenzen und wirkt sich auf nahezu alle Lebensbereiche aus. Daher wird die Lesekompetenz auch als Schlüsselkompetenz bezeichnet.

So ergaben sich z. B. bei der Weiterverfolgung der PISA 2000-Teilnehmer in Kanada Hinweise auf einen signifikanten Beitrag der Lesekompetenz im Alter von 15 Jahren zur Vorhersage des Bildungserfolgs im Alter von 19 Jahren (Knighton & Bussière, 2006). Die Analyse der Daten des *International Adult Literacy Survey* (IALS) zeigte zudem, dass Personen mit höherer Lesekompetenz bessere Chancen auf eine Beschäftigung und ein höheres Gehalt haben als Personen mit geringerer Lesekompetenz (OECD & Statistics Canada, 2000, S. 62ff.). Auch in der deutschen Arbeitswelt hat sich die Zahl der Tätigkeiten, die ohne Schriftsprachkenntnisse ausgeführt werden können, drastisch reduziert (vgl. Wölfel, Christoph, Kleinert & Heinert, 2011).

Selbst im privaten Bereich wird die Kommunikation verstärkt von persönlichen Gesprächen auf E-Mails, SMS und Chats verlegt. Die zunehmende Computerisierung macht ein gewisses Maß an Lesekompetenz selbst bei alltäglichen Angelegenheiten wie dem Fahrkartenkauf oder Bankgeschäften erforderlich (Wölfel et al., 2011). Darüber hinaus kann Lesen der Befriedigung von Unterhaltungsbedürfnissen, dem ästhetischen Erleben, der Sinnfindung und der Persönlichkeitsentfaltung dienen (Artelt et al., 2001).

Mangelnde Beherrschung der Schriftsprache bedeutet für Menschen in unserer Gesellschaft eine starke Behinderung der schulischen und akademischen Laufbahn, eine erhebliche Einschränkung beruflicher Perspektiven sowie eine Reduktion der Partizipationsmöglichkeiten am gesellschaftlichen und politischen Leben. So verwundert es nicht, dass eine geringe Lesekompetenz häufig mit emotionalen Beeinträchtigungen

[1] Das „*P*rogramme for *I*nternational *S*tudent *A*ssessment“ wird seither alle drei Jahre in fast allen 30 OECD-Ländern sowie einigen Partnerstaaten durchgeführt. Dabei werden Kompetenzen 15-Jähriger – die in Deutschland überwiegend etwa die neunte Klasse besuchen – in den Bereichen Lesen, Mathematik und Naturwissenschaften erfasst.

[2] Zugunsten der besseren Lesbarkeit wird auf die zusätzliche Nennung des weiblichen Geschlechts verzichtet. Es ist jedoch stets eingeschlossen.

und psychischen Störungen einhergeht (z. B. G. Esser & Schmidt, 1993; G. Esser, Wyschkon & Schmidt, 2002).

Dabei ist Lesen nicht als triviale Tätigkeit anzusehen. Es handelt sich vielmehr um einen sehr komplexen Prozess, der nicht einfach nur die Dekodierung eines schriftlich fixierten Bedeutungsinhaltes darstellt, sondern eine hochgradig aktive Text-Leser-Wechselwirkung, bei der der Leser unter Zuhilfenahme seines Weltwissens den Sinn eines Textes rekonstruiert. Der Erwerb von Lesekompetenz ist entsprechend mühsam und langwierig und setzt sich über das Grundschulalter hinaus fort. Inzwischen ist nicht zuletzt aufgrund der Ergebnisse verschiedener Schulleistungsstudien klar, dass auch in der Sekundarstufe gezielte Leseförderung stattfinden sollte (vgl. Hurrelmann, 2004; Streblow & Möller, 2010; Artelt et al., 2001; Gold, 2009).

Die Grundlage für gezielte Förderung bildet eine fundierte Diagnostik. Einige Studien zeigten, dass Lehrkräfte in der Sekundarstufe zum Teil Probleme haben, die Lesekompetenz ihrer Schüler adäquat einzuschätzen, was das Ergreifen angemessener Fördermaßnahmen erschweren dürfte (z. B. Karing, Matthäi & Artelt, 2011; Lorenz & Artelt, 2009; D. H. Rost & Buch, 2010). Standardisierte Lesetests können Lehrkräfte – und auch z. B. Schulpsychologen und Forscher – dabei unterstützen, das Leseverständnis zuverlässig, ökonomisch und objektiv zu überprüfen sowie gegebenenfalls adäquate Fördermaßnahmen auszuwählen und Interventionen zu evaluieren.

Lange existierten kaum Lesetests für mittlere und höhere Klassenstufen (vgl. Schneider, 2008). Dies wohl aufgrund der Annahme, dass Schüler, die die Sekundarstufe erreichen, bereits über eine ausreichende Lesekompetenz verfügen, um Texte verstehend zu lesen und aus ihnen lernen zu können. Screeningverfahren, die im unteren Leistungsbereich differenzieren, sowie Tests zur Überprüfung basaler Lesekompetenzen und des Wissens über Lesestrategien sind inzwischen auch für die Sekundarstufe erhältlich. Im Gegensatz dazu fehlte jedoch bislang ein Test, der die basale Lesekompetenz *und* das Textverständnis einschließlich höherer Verständnisebenen erfasst, der also auch im oberen Leistungsbereich zu differenzieren vermag, und der zugleich ökonomisch als Gruppentest einsetzbar ist. Aus diesem Grund wurden im Rahmen des in der vorliegenden Arbeit beschriebenen Projektes „LESEN – *L*esen *e*rmöglicht *S*inn*en*tnahme“ zwei derartige Lesetests entwickelt und evaluiert. Es handelt sich um zwei analog aufgebaute Tests: LESEN 6-7 für die Klassenstufen sechs und sieben und LESEN 8-9 für die Klassenstufen acht und neun.

Gliederung der Arbeit. Die vorliegende Arbeit gliedert sich in drei Teile: Teil I beschäftigt sich mit theoretischen Grundlagen und empirischen Befunden zum Leseverständnis und zur Leseverständnisdiagnostik in der Sekundarstufe, Teil II beschreibt die auf dem ersten Teil basierende Konstruktion und empirische Erprobung von LESEN 6-7 und LESEN 8-9 und Teil III enthält eine abschließende, kritische Diskussion der empirischen Arbeit im Hinblick auf die theoretischen Grundlagen.

In *Teil I* werden nach generellen Vorbemerkungen in Kapitel 1 in Kapitel 2 zunächst einige wichtige Begrifflichkeiten geklärt. Kapitel 3 beschäftigt sich mit dem Lesekompetenzerwerb. Dabei werden die Lesesozialisation und der Entwicklungsverlauf über die Schuljahre hinweg ebenso betrachtet wie die einzelnen Prozesse des Lesenlernens und die langfristige Bedeutung wichtiger Vorläuferkompetenzen. Kapitel 4 behandelt die einzelnen Komponenten von Lesekompetenz, wobei zunächst grob zwischen basaler Lesekompetenz und Textverständnis differenziert wird. Bei detaillierterer Betrachtung zeigt sich schließlich eine Vielzahl an (Teil-)Prozessen, woraus sich sodann die Frage nach der Dimensionalität des Konstruktes „Leseverständnis" ergibt. Bezüglich der Dimensionalität wird in einem ersten Schritt darauf eingegangen, wie viele distinkte Dimensionen sich beim Leseverständnis abgrenzen lassen, und in einem zweiten Schritt der Zusammenhang von Lese- und Hörverständnis betrachtet. Die Dimensionalität spielt auch für die in Kapitel 5 behandelte Diagnostik eine wichtige Rolle. Kapitel 5 geht dabei kurz auf mögliche klinische Diagnosen im Zusammenhang mit Leseverständnis ein und betrachtet sodann ausführlicher die diagnostische Kompetenz von Lehrkräften, insbesondere in der Sekundarstufe. Anschließend werden theoretische und methodische Grundlagen sowie Möglichkeiten und Grenzen standardisierter Lesetests diskutiert und einige für die Sekundarstufe verfügbare Tests sowie die Konzeption der PISA-Lesetests vorgestellt. Kapitel 6 fasst daraufhin aktuelle Befunde zur Lesekompetenz deutscher Sekundarschüler zusammen. Dabei wird die Leseleistung zunächst kriteriumsorientiert bzw. im Hinblick auf definierte Niveaustufen betrachtet. Anschließend wird auf Leistungsunterschiede zwischen den Bundesländern, zwischen den Schularten, zwischen den Geschlechtern sowie zwischen Schülern mit und ohne Migrationshintergrund eingegangen. Kapitel 7 behandelt schließlich verschiedene Einflussfaktoren, die die gefundenen interindividuellen und intraindividuellen Unterschiede im Hinblick auf die Lesekompetenz bewirken. Dabei finden leserseitige Einflüsse ebenso Beachtung wie Einflüsse des sozialen Umfeldes und des Textmaterials. Abhängig davon, welche Faktoren für ein Leseverständnisdefizit verantwortlich sind und welche Komponenten bzw. Prozesse betroffen sind, sollten entsprechende Fördermaßnahmen ergriffen werden. Kapitel 8 gibt daher einen kurzen Einblick in das Thema Leseförderung und zeigt mögliche Ansatzpunkte für Fördermaßnahmen auf. Ziel dieses Kapitels ist es, deutlich zu machen, welche Erkenntnisse sinnvollerweise aus einer Lesediagnostik hervorgehen sollten, damit darauf aufbauend adäquate Fördermaßnahmen ausgewählt werden können. Kapitel 9 fasst schließlich die im ersten Teil der Arbeit beschriebenen Theorien und Befunde zusammen, wobei der Fokus auf den Implikationen liegt, die sich daraus für die in Teil II beschriebene Testkonstruktion ergeben.

Teil II beginnt mit Kapitel 10, welches die Zielsetzung der Konstruktion von LESEN 6-7 und LESEN 8-9 beschreibt. Auf einen Überblick über das Projekt in Kapitel 11 und allgemeine Vorbemerkungen in Kapitel 12 folgt in Kapitel 13 eine detaillierte Darstellung der Testkonstruktion. Kapitel 13 umfasst dabei neben der theoretischen Fundie-

rung und ersten Testentwürfen auch Voruntersuchungen und Itemselektionen sowie eine Beschreibung der Endversionen. Das anschließende Kapitel 14 behandelt die Normierung von LESEN 6-7 und LESEN 8-9, wobei die Normstichprobe auf ihre Größe, Zusammensetzung und Repräsentativität hin bewertet wird. Weiter werden die Testergebnisse deskriptiv analysiert, die Rohwertverteilungen betrachtet und die Erstellung der Normtabellen dargestellt. Kapitel 15 beschreibt auf der Grundlage der bei der Normstichprobe erhobenen Daten durchgeführte Itemanalysen sowie eine Prüfung auf Modellkonformität. Es folgen eine Darstellung umfassender Reliabilitätsanalysen in Kapitel 16 und Validitätsanalysen in Kapitel 17 sowie schließlich eine Zusammenfassung von Teil II in Kapitel 18.

Teil III enthält eine abschließende Gesamtdiskussion, welche in Kapitel 19 LESEN 6-7 und LESEN 8-9 anhand von Testgütekriterien bewertet und in Kapitel 20 auf einige Kritikpunkte sowie auf daraus ableitbare Ansatzpunkte für künftige Forschungsbemühungen näher eingeht. Ein Fazit in Kapitel 21 fasst abschließend Stärken und Schwächen von LESEN 6-7 und LESEN 8-9 zusammen.

Teil I

Theoretische Grundlagen und empirische Befunde zum Leseverständnis in der Sekundarstufe

Teil I klärt wichtige Begrifflichkeiten für die vorliegende Arbeit und stellt theoretische Grundlagen sowie empirische Befunde dar, auf welchen die in Teil II beschriebene Testkonstruktion basiert. Im Einzelnen beschreibt Teil I zunächst die wichtigsten Aspekte des Lesekompetenzerwerbs und geht anschließend auf die vielfältigen Komponenten und (Teil-)Prozesse des Lesens sowie den aktuellen Forschungsstand zur Dimensionalität des Konstrukts Leseverständnis ein. Sehr ausführlich wird sodann die Diagnostik von Leseverständnis – insbesondere in der Sekundarstufe – behandelt. Darüber hinaus werden aktuelle Befunde zur Lesekompetenz deutscher Sekundarschüler und schließlich verschiedene Einflussfaktoren auf das Leseverständnis dargestellt. Es folgt ein kurzer Einblick in Möglichkeiten der Förderung von Leseverständnis in mittleren und höheren Klassenstufen. Abschließend wird der erste Teil zusammengefasst, wobei bereits daraus resultierende Implikationen für den zweiten Teil berücksichtigt werden. Begonnen wird zunächst mit einigen generellen Vorbemerkungen und Begriffsklärungen.

Kapitel 1

Vorbemerkungen

Bevor Theorien und empirische Befunde dargestellt werden, die die Grundlage für die Lesetests LESEN 6-7 und LESEN 8-9 bilden, seien noch einige generelle Überlegungen zur Übertragbarkeit von Befunden aus anderen Sprachräumen auf die deutschen Verhältnisse, zur Vielzahl an Theorien und Befunden zur Lesekompetenz sowie zur Begriffsvielfalt im Zusammenhang mit dem Lesen vorangestellt.

Übertragbarkeit von Befunden aus anderen Sprachräumen. Bei der Betrachtung von Theorien und Befunden zum Thema Leseverständnis ist stets zu bedenken, dass dieses ein Stück weit sprachgebunden ist und sich somit nicht alle Forschungsergebnisse automatisch auf andere Sprachen übertragen lassen. Ein Großteil der Theorien und Befunde zum Leseverständnis stammt aus dem angloamerikanischen Sprachraum, daher werden an dieser Stelle die wichtigsten Unterschiede zwischen der englischen und der deutschen Schriftsprache kurz dargestellt: Im Deutschen ist überwiegend einem Graphem (kleinste bedeutungsunterscheidende Einheit im Schriftsystem) ein Phonem (kleinste bedeutungsunterscheidende sprachliche Einheit) zugeordnet, weshalb ein neues Wort in der Regel alleine aufgrund der Buchstabensequenz korrekt gelesen werden kann. Im Englischen ist die Graphem-Phonem-Zuordnung inkonsistenter. Es gibt für einen Buchstaben oder eine Buchstabensequenz häufig unterschiedliche Aussprachemöglichkeiten, z. B. wird das /a/ in dem Wort „fall“ anders ausgesprochen als in dem Wort „hat“ und wieder anders in dem Wort „park“. Beim Lesen der englischen Sprache ist der Kontext der Grapheme also von größerer Bedeutung, da die Aussprache der Vokale zum Teil von nachfolgenden Graphemen abhängt (Barth, 1999, S. 24). Eine besondere Schwierigkeit im Deutschen stellen dagegen zusammengesetzte, sehr lange Wörter dar.

Die Generalisierbarkeit englischer Modelle auf die deutschen Verhältnisse wird vor allem beim Schriftspracherwerb angezweifelt (z. B. Wimmer, Hartl & Moser, 1990). Aufgrund der Unterschiede bezüglich der Graphem-Phonem-Korrespondenzen wird die Übersetzung der graphischen bzw. orthographischen Information in eine phonologische Repräsentation im Deutschen meist als leichter erachtet als im Englischen. Für Leseprozesse auf Wort-, Satz- und Textebene, die über diese Übersetzung hinausgehen, wird hingegen ebenso wie für sozial-emotionale und motivationale Prozesse beim Lesen davon ausgegangen, dass sie im Deutschen und im Englischen vergleichbar sind (vgl. Schreier, 2004). Da sich die vorliegende Arbeit in erster Linie mit dem Lese-

verständnis in der Sekundarstufe beschäftigt und dort Leseprozesse im Vordergrund stehen, die auf höheren Verständnisebenen angesiedelt sind, spielen Unterschiede zwischen den Sprachräumen eine geringere Rolle.

Vielzahl an Theorien und Befunden. Weiter ist vorab anzumerken, dass in den letzten zwei Jahrzehnten, vor allem infolge der ersten PISA-Befunde, die Zahl der Publikationen zum Thema Lesekompetenz enorm angestiegen ist. Inzwischen existieren so viele Theorien, Modelle und Studien zur Lesekompetenz, dass dieses Feld kaum noch zu überblicken ist. Allein die Anzahl der Publikationen zu jeder einzelnen PISA-Erhebung ist sehr groß. Hinzu kommen andere große nationale und internationale Schulleistungsstudien sowie kleinere Forschungsprojekte, die sich mit der Thematik des Lesens befassen. Allein für das Jahr 2012 findet die Psychologiedatenbank PsychINFO für den Schlüsselbegriff „reading comprehension“ 245 Treffer in Form von Zeitschriftenartikeln, Dissertationen und Büchern bzw. Buchbeiträgen. Bei Google Scholar finden sich bei der Auswahl „seit 2012“ 29 500 Einträge zu „reading comprehension“ (Stand am 16.10.2013).

Aus diesem Grund wird in der vorliegenden Arbeit kein Anspruch auf vollständige Berücksichtigung aller publizierten Theorien und Befunde erhoben; vielmehr werden die bedeutendsten und aktuellsten Publikationen zum Leseverständnis und zur Leseverständnisdiagnostik im Sekundarschulalter berücksichtigt, da die Konstruktion von LESEN 6-7 und LESEN 8-9 auf diesen basiert.

Begriffsvielfalt. Aufgrund der Komplexität des Leseprozesses und der großen Anzahl an Forschern und Praktikern, die sich damit beschäftigen, ist zudem eine unübersichtliche Begriffsvielfalt entstanden. Die zusätzliche Verwendung von und Vermischung mit Begriffen aus dem englischen Sprachraum führt zu weiterer Verkomplizierung. Häufig bleibt unklar, was genau die Autoren z. B. unter Lesekompetenz, Leseverständnis oder Reading Literacy verstehen. Zum einen werden diese Begriffe häufig synonym verwendet, zum anderen wird ein und derselbe Begriff in unterschiedlicher Bedeutung verwendet oder verschiedene Konzeptionen werden sogar vermischt. Dies erschwert Vergleiche von Ergebnissen verschiedener Studien. Daher werden im folgenden Kapitel zunächst einige Begrifflichkeiten geklärt.

Kapitel 2

Begrifflichkeiten

Die Begriffe „Lesekompetenz“, „Reading Literacy“ (bzw. „Literalität“) und „Leseverständnis“ sind eng verwandt und werden häufig synonym verwendet. Die Hauptunterschiede liegen in den unterschiedlichen theoretischen Traditionen, denen sie entstammen, sowie in der Breite ihrer Definition (W. Lenhard, 2013, S. 45f.). Da die Begriffe in verschiedenen Disziplinen in unterschiedlicher Bedeutung verwendet werden, wird für die nachfolgende Abgrenzung keine Allgemeingültigkeit beansprucht. Am Ende jeder Begriffsklärung wird festgelegt, in welcher Bedeutung der jeweilige Begriff in der vorliegenden Arbeit verwendet wird.

Lesekompetenz. Der Kompetenzbegriff im Allgemeinen wird sowohl in der öffentlichen Diskussion als auch in der Wissenschaft sehr unterschiedlich gebraucht. Die Bandbreite reicht einerseits vom angeborenen Persönlichkeitsmerkmal bis zum erworbenen Wissensschatz und andererseits von der generellen Schlüsselqualifikation bis zur spezifischen Fertigkeit (vgl. Artelt et al., 2007). In der pädagogischen Psychologie und der empirischen Bildungsforschung entwickelte sich der Begriff „Kompetenz“ aus dem Kontext der Expertiseforschung und wird inzwischen nahezu inflationär gebraucht (W. Lenhard, 2013, S. 47f.). In seiner ursprünglichen Bedeutung bezeichnet er die Leistung einer Person, die aus einem Zusammenwirken von Fähigkeit, Fertigkeit, Vorwissen, Erfahrung und Motivation resultiert.

*Lese*kompetenz im Speziellen kann entsprechend als Leseverständnisleistung angesehen werden, die über kognitive Faktoren hinaus auch motivationale und affektive Aspekte einschließt (W. Lenhard, 2013, S. 46f.). Die motivationalen und affektiven Aspekte sind entscheidend dafür, ob ein Potenzial in einer Anwendungssituation tatsächlich ausgeschöpft wird. Einigen didaktischen sowie literatur- und kulturwissenschaftlichen Modellen zufolge sind über die genannten leserseitigen Faktoren hinaus auch soziale und textseitige Faktoren entscheidend für die Lesekompetenz (s. z. B. Hurrelmann, 2006). W. Lenhard (2013, S. 46) differenziert entsprechend zwischen der *Lesekompetenz im engeren Sinne*, also der Leseverständnisleistung einschließlich motivationaler und affektiver Faktoren, und der *Lesekompetenz im weiteren Sinne*, welche darüber hinaus auch soziale Faktoren und die Interaktion des Lesers mit dem Text umfasst. Im Rahmen der Definition im weiteren Sinne ist Lesekompetenz somit keine individuelle Eigenschaft, sondern variiert auch abhängig von externen Faktoren, da sie nicht nur von der individuellen Fähigkeit und Motivation, sondern auch vom sozialen Kontext und den Eigenschaften des Textes beeinflusst wird. Diese erweiterte Betrach-

tung von Lesekompetenz klingt plausibel und mag aus didaktischer Sicht sinnvoll sein, jedoch ist sie schwer fassbar. Daher existieren auch keine entsprechenden Messinstrumente, was wiederum zur Folge hat, dass die Lesekompetenz im weiteren Sinne der empirischen Forschung derzeit kaum zugänglich ist (vgl. W. Lenhard, 2013; Wild & Möller, 2009).

Im Zuge der internationalen Schulleistungsvergleichsstudien hat sich der Gebrauch des Kompetenzbegriffs in den letzten Jahrzehnten jedoch verändert. Er wird heute häufig nicht mehr nur für die erbrachte Leistung, sondern auch für ein Leistungspotenzial im Sinne einer Fähigkeit verwendet (vgl. W. Lenhard, 2013, S. 47f.). Meistzitiert dürfte inzwischen die PISA-Definition sein, der zufolge Lesekompetenz die Fähigkeit darstellt, „geschriebene Texte zu verstehen, zu nutzen und über sie zu reflektieren, um eigene Ziele zu erreichen, das eigene Wissen und Potenzial weiterzuentwickeln und am gesellschaftlichen Leben teilzunehmen“ (Baumert, Stanat & Demmrich, 2001, S. 23). Auch die in dieser Definition deutlich werdende stärkere Orientierung des Kompetenzbegriffs am angloamerikanischen Literacy-Konzept, welches nachfolgend beschrieben wird, ist wohl auf die internationalen Schulleistungsvergleichsstudien zurückzuführen.

In der vorliegenden Arbeit wird der Begriff Lesekompetenz gemäß der Definition „Lesekompetenz im engeren Sinn“ von W. Lenhard (2013, S. 46ff.) verwendet. D. h. der Begriff bezieht sich auf die Leseverständnis*leistung*, wodurch über die Fähigkeit zum Textverständnis hinaus motivationale und affektive Faktoren eingeschlossen sind. Wird jedoch von PISA-Ergebnissen zur Lesekompetenz berichtet, wurden diese mit einem Test erhoben, dem die oben zitierte Definition von Baumert et al. (2001, S. 23) zugrunde liegt. Diese ist breiter angelegt als in der vorliegenden Arbeit. Die PISA-Lesetests werden in Kapitel 5.3.5 genauer beschrieben, sodass noch deutlicher werden wird, was sie erfassen. Weiter ist einschränkend anzumerken, dass nicht aus allen zitierten Studien eindeutig hervorgeht, was jeweils unter Lesekompetenz verstanden wird, sodass sich eine eindeutige und konsequente Verwendung des Begriffs nicht einhalten lässt.

Reading Literacy und Literalität. Der Begriff „Literacy“ ist schon aus der Antike bekannt und war noch nie eindeutig definiert (Gough, Hoover & Peterson, 1996). Zum einen wurde darunter die Fähigkeit zu lesen und zu schreiben verstanden, zum anderen wurde der Begriff synonym mit „kultiviert“ und „gebildet“ verwendet. Inzwischen umfasst Literacy noch weitere Bereiche, z. B. spricht man von „music literacy“, „computer literacy“ oder „visual literacy“. Der Begriff geht somit über das klassische Lesen von gedrucktem Text oder die literarische Bildung hinaus und bezieht sich auf die kompetente Nutzung verschiedener Medien (vgl. Bachmair, 2009, S. 26). Die Fähigkeit zu lesen und zu schreiben muss gar nicht mehr impliziert sein (Gough et al., 1996). Literacy umfasst generell Basiskompetenzen, die zur befriedigenden Lebensführung in der modernen Gesellschaft notwendig sind (Artelt et al., 2001).

Reading Literacy bezieht sich dagegen explizit auf das Lesen. Sie ist funktional orientiert und wird als notwendiges Kulturwerkzeug verstanden, um in der modernen Informations- und Kommunikationsgesellschaft bestehen zu können (Artelt et al., 2001). Als fächerübergreifende Schlüsselkompetenz beinhaltet sie das Verstehen und Nutzen von geschriebenen Texten sowie das Reflektieren über Gelesenes. Sie dient dazu, bestimmte Ziele zu erreichen, eigene Kenntnisse und Fähigkeiten zu erweitern und am gesellschaftlichen Leben zu partizipieren (Baumert et al., 2001). Somit verschob sich der Schwerpunkt des Lesekompetenzkonzepts im Zuge der stärkeren Fokussierung auf die Reading Literacy von der traditionellen mitteleuropäischen sprachlich-literarischen Grundbildung auf die Rezeption und Verarbeitung von Informationstexten (vgl. Groeben, 2004; Hurrelmann, 2006). Im Deutschen wird für Reading Literacy häufig auch der Begriff „Literalität“ verwendet.

„Reading Literacy“ ist also noch breiter gefasst als der Begriff der Lesekompetenz. In der vorliegenden Arbeit werden die Begriffe „Reading Literacy“ und „Literalität“ gemäß der PISA-Definition (Baumert et al., 2001, S. 23) bzw. der Definition von W. Lenhard (2013, S. 46) verwendet. Sie beschreiben somit die *Fähigkeit, Lesen in alltäglichen Anwendungssituationen und im kulturellen Kontext einsetzen zu können*.

Leseverständnis. Leseverständnis stellt das im Vergleich zu den zuvor erläuterten Begriffen am engsten gefasste Konstrukt dar. Es bezeichnet die *Fähigkeit, Textinhalte zu rekonstruieren und zugleich die dazu notwendigen kognitiven Prozesse* (W. Lenhard, 2013, S. 46). Dazu zählen sowohl basale Lesekompetenzen im Sinne der technischen Lesefertigkeit als auch die Textverarbeitung auf höheren Verständnisebenen, einschließlich des Aufbaus einer mentalen Repräsentation des im Text dargestellten Sachverhalts. Hier geht es nicht um tatsächlich erbrachte Leistung, sondern um die prinzipielle Fähigkeit bzw. das Potenzial, Texten Sinn zu entnehmen – unabhängig von der aktuellen Motivation oder Bereitschaft (W. Lenhard, 2013, S. 46f.). Mit dem Leseverständnis beschäftigen sich vor allem die Kognitionspsychologie und die experimentelle Leseforschung.

In der vorliegenden Arbeit bezeichnet der Begriff Leseverständnis gemäß W. Lenhard (2013, S. 46) die Fähigkeit, schriftlich fixierte Bedeutungsinhalte zu rekonstruieren, und zugleich die dabei ablaufenden kognitiven Prozesse.

Insgesamt werden die Begriffe „Leseverständnis“, „Lesekompetenz“ und „Reading Literacy“ somit in der vorliegenden Arbeit folgendermaßen verstanden (s. auch Abb. 1): *Leseverständnis* umfasst das prinzipielle Potenzial sowie die kognitiven Prozesse zur Rekonstruktion eines Textinhaltes und bildet den Kern der Lesekompetenz. *Lesekompetenz* erweitert den Leseverständnisbegriff um motivationale und affektive Faktoren und bezeichnet die in einer Anwendungssituation tatsächlich erbrachte Leistung. Das umfassendste der drei in diesem Kapitel erläuterten Konstrukte stellt die *Reading Literacy* dar, die Leseverständnis und Lesekompetenz beinhaltet und darüber hinaus den Alltagsbezug sowie den kulturellen Kontext einschließt.

Die vorliegende Arbeit beschäftigt sich vorwiegend mit dem Leseverständnis und der Lesekompetenz sowie deren Messung in der Sekundarstufe. Entsprechend liegt der Schwerpunkt auf allgemeinpsychologisch-kognitionspsychologischen und psychometrischen Aspekten des Lesens. Im Folgenden wird das Lesen jedoch zunächst aus der Entwicklungsperspektive betrachtet.

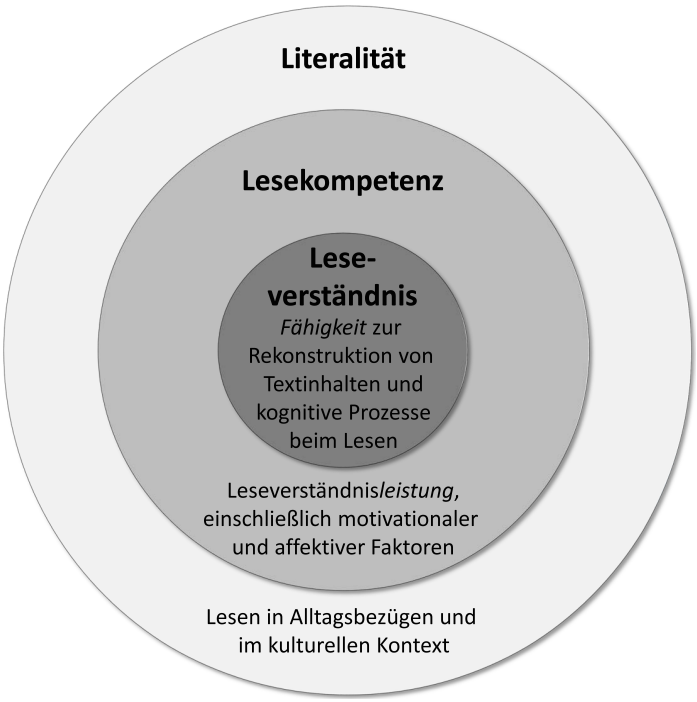


Abbildung 1. Relation der Begriffe Leseverständnis, Lesekompetenz und Literalität und ihre Verwendung in der vorliegenden Arbeit (in Anlehnung an W. Lenhard, 2013, S. 46).

Kapitel 3

Entwicklung von Lesekompetenz

Lesenlernen ist ein komplexer Prozess und der Weg vom ersten Interesse für Buchstaben und vorgelesene Geschichten bis zum selbstständigen und zielgerichteten Nutzen schriftlicher Texte lang (Artelt et al., 2001). Das vorliegende Kapitel betrachtet zunächst die wichtigsten Aspekte von Lesesozialisation und vom Entwicklungsverlauf des Lesekompetenzerwerbs. Anschließend werden am Lesenlernen beteiligte Prozesse sowie langfristige Auswirkungen früher Vorläuferkompetenzen des Lesens dargestellt. Aufgrund der Relevanz für die vorliegende Arbeit finden dabei jeweils vor allem diejenigen Lernprozesse Berücksichtigung, die sich üblicherweise im Sekundarschulalter vollziehen bzw. die für die Lesekompetenz in mittleren und höheren Klassenstufen besonders wichtig sind.

3.1 Lesesozialisation und Entwicklungsverlauf

Als Lesesozialisation bezeichnet man „alle sozial und individuell bedingten Prozesse, die im Verlauf des Lebens dazu führen, dass Menschen (nicht) die Fähigkeiten, die Motivation und die Praxis entwickeln, schriftsprachliche [...] Texte [...] zu rezipieren“ (Philipp, 2011b, S. 29). Lesesozialisation ist ein interindividuell sehr unterschiedlich verlaufender Prozess, der von Merkmalen wie Geschlecht, soziale Herkunft und Anregung durch die Umwelt sowie von der Interaktion dieser Faktoren abhängt (Philipp, 2011b, S. 22).

Schon lange vor Schuleintritt sammeln Kinder erste Erfahrungen mit Schriftsprache, Texten und literarischen Formen, wenn ihnen z. B. Geschichten erzählt oder Bilderbücher vorgelesen werden (Rosebrock & Nix, 2011; Artelt et al., 2007). Zudem wird ihnen Wissen über Funktionen des Lesens vermittelt und Anschlusskommunikation geübt (Purcell-Gates, 1989). Auch metasprachliche Kompetenzen, Wertschätzung des Lesens und Motivation zum Lesen werden durch Kommunikation und Vorleben in der Familie weitergegeben (Artelt et al., 2007). Die Anregung und Unterstützung, welche Kinder diesbezüglich in ihren Familien erfahren, variieren sowohl in qualitativer als auch in quantitativer Hinsicht.

Im Vorschulalter werden dann wichtige Vorläuferfertigkeiten für das Lesen erworben (Rosebrock & Nix, 2011; Artelt et al., 2007). Hier findet teilweise in der Vorschule

eine systematische Förderung z. B. phonologischer Kompetenzen statt (z. B. W. Lenhard, 2013, S. 117f.).

In der Schule erfolgt schließlich die Alphabetisierung im Sinne einer methodisch geplanten, gegenstandsbezogenen und zielorientierten Vermittlung von Wissen, Fertigkeiten und kulturellen Werten (Hurrelmann, 2004). Abhängig von vorherigen Bedingungen der Lesesozialisation sehen sich Lehrkräfte dabei teilweise einer sehr heterogenen Gruppe von Lesern gegenüber (Philipp, 2011b, S. 22). Im Verlauf eines Grundschuljahres zeigen sich im Hinblick auf das Leseverständnis deutliche Leistungszuwächse. Im deutschen Sprachraum beherrscht ein Großteil der Kinder bereits am Ende der ersten Klasse die Graphem-Phonem-Zuordnung und die Lautsynthese (vgl. Artelt et al., 2007; Schneider, 2008). Der Leistungszuwachs im Leseverständnis von der ersten zur zweiten Klassenstufe beträgt etwa eine Standardabweichung (vgl. Philipp, 2011b, S. 74). Schon hier bestehen allerdings große interindividuelle Differenzen in der Lesegeschwindigkeit, die bis zum Ende der Grundschulzeit bestehen bleiben (Klicpera, Humer, Lugmayr & Gasteiger Klicpera, 1993; Schneider & Näslund, 1999). Nach der ersten Klassenstufe nehmen die Leistungszuwächse im Leseverständnis pro Schuljahr stetig ab (vgl. Philipp, 2011b, S. 74). Von der dritten zur vierten Klassenstufe liegen diese nur noch bei einer drittel bis einer halben Standardabweichung.

Personen, die sich später als gute Leser erweisen, erreichen in der späten Kindheit häufig eine Viellesephase, was die automatische Worterkennung und die Leseflüssigkeit weiter fördert (Rosebrock & Nix, 2011, S. 21). Beim Übertritt in die Sekundarstufe wird erwartet, dass ein Übergang vom „learning-to-read“ zum „reading-to-learn“ möglich ist. Am Ende der Kindheit sollte die technische Lesefertigkeit dann so weit entwickelt sein, dass kognitive Ressourcen für inhaltliche Textverarbeitung frei werden und aus anspruchsvolleren Texten gelernt werden kann.

In der Jugend sinkt der familiäre und schulische Einfluss auf das Leseverhalten, während der Einfluss der Gleichaltrigen zunimmt, z. B. bei Gesprächen über Lektürewahl, Textinterpretation und -bewertung (Artelt et al., 2007). Auch in der Sekundarstufe finden sich weiterhin Leistungszuwächse im Verlauf der einzelnen Schuljahre (W. Lenhard, 2013, S. 40). Diese fallen jedoch geringer aus als in der Grundschule und nehmen zudem über die Sekundarschuljahre hinweg weiter ab. Einige Studien fanden in den ersten Sekundarschuljahren Leistungszuwächse in der Größenordnung von etwa einer drittel bis einer halben Standardabweichung pro Schuljahr (z. B. Retelsdorf & Möller, 2008; Baumert & Artelt, 2002; R. Lehmann & Lenkeit, 2008). Zumindest zum Beginn der Sekundarschulzeit scheinen die Leistungszuwächse bei Schülern der verschiedenen Regelschularten recht parallel zu verlaufen, ebenso wie die Leistungen von Schülern mit und ohne Migrationshintergrund gleichermaßen anzusteigen scheinen (Retelsdorf & Möller, 2008; Gräsel, Göbel & Stark, 2007). Es findet sich weder ein Aufholen der jeweils schwächeren Gruppe (Hauptschüler bzw. Schüler mit Migrationshintergrund) noch ein Schereneffekt im Sinne einer zunehmenden Leistungsdiskrepanz. Im weiteren Verlauf der Sekundarschulzeit gibt es jedoch Hinweise darauf, dass es zu-

mindest beim Vergleich von Gymnasiasten mit Nicht-Gymnasiasten doch noch zu einem Schereneffekt im Hinblick auf die Dekodiergeschwindigkeit kommt, während die Leistungszuwächse im Leseverständnis weiterhin recht parallel verlaufen (Retelsdorf & Möller, 2011). In der DESI-Studie[3] fiel der Anstieg der Lesekompetenz vom Beginn zum Ende der neunten Klasse nicht signifikant aus (DESI-Konsortium, 2006, S. 6f.). Selbst bei einer separierten Betrachtung der verschiedenen Bildungsgänge fanden sich nur geringfügige Veränderungen hinsichtlich der Verteilung der Schüler über verschiedene Stufen der Lesekompetenz. Lediglich für den gymnasialen Bildungsgang ließen sich die Zuwächse in der Lesekompetenz zufallskritisch absichern. Selbst dort erwies sich der Anstieg des Mittelwerts jedoch als eher gering. In Klassen, die verstärkt mit literarischen Texten arbeiteten, war allerdings durchaus ein deutlicherer Kompetenzanstieg über das neunte Schuljahr hinweg messbar.

In höheren Klassenstufen fallen die Leistungszuwächse dann generell noch geringer aus. Beispielsweise nimmt die Leseverständnisleistung von der neunten zur zehnten Jahrgangsstufe und von der zehnten zur elften Jahrgangsstufe nur noch zwischen 0.1 und 0.2 Standardabweichungen zu (Philipp, 2011b, S. 74).

Der individuelle Entwicklungsverlauf hängt insgesamt stark von den Bedingungen der Lesesozialisation des einzelnen Kindes ab. Ein idealtypischer Verlauf findet sich hauptsächlich bei Mädchen der Mittelschicht, nicht jedoch z. B. bei Jungen der Unterschicht (Rosebrock, 2004). In der Realität hatte ein Viertel der Kinder schon im Vorschulalter wenig Kontakt mit schriftlichen Medien und beteiligt sich nicht aktiv an der Leseentwicklung. Diese Schüler versuchen Lesesituationen zu vermeiden. Es entsteht ein Teufelskreis aus fehlender Motivation und mangelnder Kompetenz. Nur etwa ein Drittel der Schüler liest in der Pubertät in der Freizeit. Für problemloses Lernen aus Texten reicht die basale Lesefertigkeit in der Realität nicht einmal in den fünften Klassen der Gymnasien bei allen Schülern aus. Selbst in mittleren Klassenstufen der Sekundarstufe zeigen sich teilweise noch grundlegende Schwierigkeiten im Umgang mit Texten.

3.2 Prozesse des Lesenlernens

Es existieren zahlreiche Modelle zu den am Schriftspracherwerb beteiligten Prozessen. Die meisten dieser Modelle konzentrieren sich auf Prozesse des Buchstaben- und Wortlesens. Nachfolgende Darstellung orientiert sich daher im wesentlichen am Prozessmodell des Lesenlernens von H. Marx und Jungmann (2000), das verschiedene Modelle der Lesekompetenz und des Lesekompetenzerwerbs integriert und auch das Textverstehen einschließt. Es berücksichtigt die vorschulischen und schulischen Lese-

[3] Die Studie *D*eutsch-*E*nglisch-*S*chülerleistungen-*I*nternational (DESI) untersuchte im Schuljahr 2003/2004 die sprachlichen Leistungen von Schülern in der neunten Klassenstufe in den Fächern Deutsch und Englisch und orientierte sich an den Curricula der neunten Klassenstufe.

lernprozesse und integriert z. B. den „Simple View of Reading“ (z. B. Gough & Tunmer, 1986, s. auch Kap. 5), konnektionistische Leselernmodelle (z. B. M. J. Adams, 1990; M. Seidenberg & McClelland, 1989; Van Orden, Pennington & Stone, 1990) und Stufenmodelle (z. B. Frith, 1985; Günther, 1986).

Das Prozessmodell von H. Marx und Jungmann (2000) geht davon aus, dass beim Lesenlernen vorhandene unspezifische Vorläuferfertigkeiten und neu zu erwerbende, lesespezifische Fertigkeiten integriert werden müssen. Als unspezifische Vorläuferfertigkeiten, die auch für andere – z. B. mathematische oder musikalische – Erwerbsprozesse nötig sind, werden intellektuelle Fähigkeiten, Arbeitsgedächtnisleistungen und mündlicher Sprachgebrauch angesehen. Sie sind für den Schriftspracherwerb notwendig, aber nicht hinreichend. Zu den lesespezifischen Fertigkeiten gehören phonologische Informationsverarbeitungskompetenzen, Buchstabenkenntnis, die Übersetzung graphischer bzw. orthographischer Informationen in eine phonologische Repräsentation sowie der Abruf der entsprechenden Wortbedeutung. Die lesespezifischen Fertigkeiten werden im Idealfall in der genannten Reihenfolge erworben (H. Marx & Reinhold, 2010). Im nachfolgenden Kapitel wird noch einmal detailliert auf die Bedeutung der frühen phonologischen Informationsverarbeitungskompetenzen sowie der generellen sprachlichen Fähigkeiten eingegangen.

Mit zunehmender Geübtheit im Lesen entsteht schließlich implizites Wissen darüber, welche phonologische Umsetzung für bestimmte Graphemgruppen am wahrscheinlichsten korrekt ist (Treiman, Kessler, Zevin, Bick & Davis, 2006; H. Marx & Reinhold, 2010). Dieses Wissen erleichtert den Lese- und Verstehensprozess, und es wird selbst vom semantischen Hörverstehen auf Wortebene beeinflusst, denn der Inhalt eines geschriebenen Wortes kann nur dann richtig verstanden werden, wenn dieses Wort auch im mündlichen Wortschatz vorkommt und das semantische Lexikon des Gedächtnisses zugänglich ist (H. Marx & Reinhold, 2010). Zudem erleichtert orthographisches Wissen die korrekte phonologische Realisation und die Sinnerfassung. Für das Leseverstehen auf Satz- und Textebene ist darüber hinaus Wissen über textstrukturelle Aspekte – z. B. Wortwahl, Grammatik und Syntax – sowie Vorwissen zum Textinhalt und Hörverstehen bezüglich mündlicher Mitteilungen von Bedeutung (H. Marx & Reinhold, 2010).

H. Marx und Jungmann (2000) fanden, dass die Teilfertigkeiten des Lesens zu Beginn des Lesenlernens noch recht unabhängig voneinander sind, jedoch im Laufe der Grundschuljahre, d. h. mit zunehmender Schriftspracherfahrung, zu *einer* Lesefertigkeit werden. Das Hörverstehen stellt dabei die Obergrenze des Leseverstehens dar und ist vor allem bei geübten Lesern ein wesentlicher Prädiktor des Textverstehens (vgl. auch Kap. 4.3).

Zugleich nimmt im Laufe der Schuljahre das Wissen über und der effektive Einsatz von Lern- und Behaltensstrategien stetig zu (vgl. Artelt et al., 2007). Strategien werden dabei zunächst bereichsspezifisch und dann zunehmend in verschiedenen Kontexten eingesetzt (vgl. Schneider & Pressley, 1989, S. 148ff.). Artelt et al. (2007) nehmen

an, dass sich sowohl das deklarative Wissen über adäquates Vorgehen beim Lesen als auch die prozedurale Fähigkeit zur Überwachung und Regulation während des Lesens erst im späten Kindes- und frühen Jugendalter erheblich steigern (vgl. auch Schneider & Pressley, 1989; Hasselhorn, 2010). In diesem Alter sollten basale Lesekompetenzen soweit automatisiert sein, dass kognitive Kapazitäten für die Begleitung des Leseprozesses auf der Metaebene sowie einen entsprechenden Strategieeinsatz frei werden.

3.3 Langfristige Bedeutung wichtiger Vorläuferkompetenzen

Zahlreiche nationale und internationale Studien zeigten inzwischen, dass im Vorschulalter vorhandene Kompetenzen im Bereich der phonologischen Informationsverarbeitung für den späteren Schriftspracherwerb von großer Bedeutung sind (Schneider, 2008). Zu den Kompetenzen der phonologischen Informationsverarbeitung zählen neben der phonologischen Bewusstheit, also der Einsicht in die Phonologie der Sprache, auch das phonetische Rekodieren im Arbeitsgedächtnis, also das kurzzeitige Aufrechterhalten auditiver Informationen, sowie das phonologische Rekodieren beim Zugriff auf das semantische Lexikon.

Eine Längsschnittstudie über acht Jahre zeigte, dass die vorschulische *phonologische Bewusstheit* im Deutschen die Leseleistungen in der ersten, aber nicht in höheren Klassenstufen vorhersagen kann. Die phonologische Bewusstheit erwies sich als bedeutsam für den Erwerb der Dekodierfähigkeit bei Leseanfängern, verlor jedoch an Bedeutung, sobald die Lesefertigkeiten stärker automatisiert waren und mehr Lesestrategien zur Verfügung standen (Wimmer, Mayringer & Landerl, 2000; Wimmer & Mayringer, 2002; Landerl & Wimmer, 2008). Kinder, bei denen im Vorschulalter Defizite in der phonologischen Bewusstheit vorlagen, zeigten im Verlauf der Grundschulzeit vor allem Rechtschreibschwierigkeiten, während sich die Leseleistungen oft unauffällig entwickelten.

Auch ein Zusammenhang zwischen Schriftsprachkompetenzen und dem *phonetischen Rekodieren im Arbeitsgedächtnis* ist empirisch nachgewiesen. So lässt sich aus der vorschulischen Gedächtnisspanne sowie aus der Leistung beim Nachsprechen von Kunstwörtern die spätere Schriftsprachkompetenz vorhersagen (Gathercole & Baddeley, 1993; Gathercole, Willis & Baddeley, 1991; Näslund & Schneider, 1996). Schneider (2008) nimmt an, dass die Kapazität des phonologischen Arbeitsgedächtnisses insbesondere für Leseanfänger von Bedeutung ist, da bei ihnen die Übersetzung der Grapheme in Phoneme noch relativ langsam abläuft und die einzelnen Laute länger im Arbeitsgedächtnis aufrecht erhalten werden müssen als bei geübten Lesern. Tatsächlich zeigte sich längsschnittlich, dass das phonologische Arbeitsgedächtnis lediglich die Leseflüssigkeit in der ersten Klasse vorhersagen kann, während es auf die Leseleistung in höheren Klassenstufen sowie auf die Rechtschreibleistung keinen bedeutsamen Einfluss zu haben scheint (Landerl & Wimmer, 2008).

Bezüglich des *phonologischen Rekodierens beim Zugriff auf das semantische Lexikon* ergab sich, dass bei Kindern, die im Vorschulalter einen verlangsamten Zugriff auf das Langzeitgedächtnis aufwiesen, Probleme beim Lesenlernen auftraten (Wimmer & Mayringer, 2002; Wimmer et al., 2000). Wurde die Geschwindigkeit beim Zugriff auf das Langzeitgedächtnis über das schnelle Benennen von Objekten, Farben, Buchstaben oder Ziffern überprüft, erlaubten die Leistungen im schnellen Benennen zu Beginn der ersten Klasse eine signifikante Vorhersage der Lesegeschwindigkeit bis zur achten Klasse. Die Benenngeschwindigkeit konnte einen Varianzanteil der Leseflüssigkeit aufklären, der nicht durch andere Komponenten der phonologischen Informationsverarbeitung erklärt werden konnte (Landerl & Wimmer, 2008). Zwar ist die Leseleistung relativ stabil und kann am besten durch die Leistung im vorherigen Schuljahr vorhergesagt werden, jedoch kann die vorschulische Leistung beim schnellen Benennen darüber hinaus Varianz der Leseleistung aufklären (de Jong & van der Leij, 1999). Konform mit diesen Befunden erwies sich eine geringe Benenngeschwindigkeit in vielen Sprachräumen als ein zentrales Defizit bei Kindern mit Lesestörungen (z. B. Willburger, Fussenegger, Moll, Wood & Landerl, 2008; Wolf & Bowers, 1999). Kinder, die im Vorschulalter Defizite beim Abruf aus dem Langzeitgedächtnis zeigen, entwickeln in der Regel eine geringe Lesegeschwindigkeit, während die Rechtschreibleistungen meist unauffällig sind (Wimmer & Mayringer, 2002; Wimmer et al., 2000).

Weitere Studien zeigten, dass vorschulische *Kompetenzen im Bereich der mündlichen Sprache* (Wortschatz und Grammatik) vor allem das Leseverstehen vorhersagen können, während Lesegenauigkeit und Lesegeschwindigkeit im Grundschulalter durch Maße der phonologischen Informationsverarbeitung hinreichend erklärt werden können, ohne dass breitere sprachliche Fähigkeiten zusätzlich zur Varianzaufklärung beitrügen (Catts, 1993; Catts & Hogan, 2003; Ennemoser, Marx, Weber & Schneider, 2012; Roth, Speece & Cooper, 2002). Kinder, die bei gut ausgeprägter Dekodierleistung Leseverständnisprobleme aufweisen, haben meist auch Schwierigkeiten beim Hörverstehen, d. h. es fällt ihnen beispielsweise schwer, Fragen zu gehörten Geschichten zu beantworten (Stothard & Hulme, 1992). Snowling, Bishop und Stothard (2000) fanden bei Kindern mit vorschulischen Sprachdefiziten zwischen dem Alter von 8 und 15 Jahren zunehmende Leseschwierigkeiten. Dieser Befund steht im Einklang mit der Annahme von Grimm (1995) und Befunden von Ennemoser et al. (2012), dass grammatische und linguistische Kompetenzen vor allem im späteren Verlauf der Leseentwicklung beim verstehenden Lesen an Bedeutung gewinnen. Auch H. Marx und Jungmann (2000) zeigten, dass der Zusammenhang von Sprach- und Leseverständnis mit zunehmender Leseerfahrung steigt. Insgesamt deuten somit zahlreiche Befunde darauf hin, dass sprachliche Kompetenzen für die Lesekompetenz, insbesondere für das Leseverständnis, relevant sind und bereits im Vorschulalter eine Prognose späterer Leseverständnisleistungen erlauben.

3.4 Fazit

Die Entwicklung von Lesekompetenz beginnt also bereits vor dem Schuleintritt. In der Familie, im Kindergarten und in der Vorschule werden nicht nur lesebezogene Einstellungen und Werte, sondern auch wichtige Vorläuferfertigkeiten für den Schriftspracherwerb vermittelt. Hier zeigte sich, dass die vorschulische phonologische Bewusstheit und das phonologische Arbeitsgedächtnis vor allem für das erste Jahr des Schriftspracherwerbs prädiktiv sind. Für die Langzeitentwicklung der Lesegeschwindigkeit ist dagegen die Benenngeschwindigkeit bzw. die Geschwindigkeit beim Zugriff auf das Langzeitgedächtnis bedeutsamer. Für das Leseverständnis sind vor allem sprachliche Kompetenzen (Wortschatz und Grammatik) langfristig prädiktiv.

In der Grundschulzeit wird in der Regel die Graphem-Phonem-Korrespondenz erlernt und basale Lesefertigkeiten werden trainiert. In der Sekundarstufe sollte schließlich eine für tiefergehendes Textverständnis notwendige Automatisierung und Geschwindigkeit basaler Leseprozesse gegeben sein. Jedoch zeigen empirische Befunde, dass dies nicht bei allen Sekundarschülern der Fall ist. Die Geschwindigkeit der basalen Leseprozesse nimmt vielmehr sogar noch bis in frühe Erwachsenenalter hinein zu (Schneider, 2008). Vor allem aber im Hinblick auf das Textverständnis und verstehensförderliche Lesestrategien sind auch nach der Grundschulzeit noch wichtige Entwicklungsschritte zu vollziehen. Dennoch ist der Leistungszuwachs im Leseverständnis innerhalb eines Schuljahres in der Sekundarstufe deutlich geringer als noch in der Grundschulzeit.

Im Modell von H. Marx und Reinhold (2010) wird das Lesenlernen als Entwicklungsprozess von aufmerksamkeitskontrollierter hin zu automatischer Textverarbeitung beschrieben. Leseanfänger müssen zunächst aufmerksamkeitskontrolliert die einzelnen Teilfertigkeiten und -prozesse erwerben und erreichen schließlich durch intensives Üben einen Automatisierungsgrad, der es ihnen ermöglicht, mühelos komplexere Einheiten zu verarbeiten. Dadurch wird das Arbeitsgedächtnis entlastet und Ressourcen werden freigestellt, die zum inhaltlichen Verständnis des Gelesenen genutzt werden können.

Bei der groben Skizzierung der Entwicklung von Lesekompetenz in den vergangenen Abschnitten wurden bereits einige für das Leseverständnis bedeutsame Teilprozesse angesprochen. Im nun folgenden Kapitel werden die verschiedenen Teilprozesse und Komponenten von Leseverständnis nun im Detail betrachtet.

Kapitel 4

Komponenten von Leseverständnis

Zu den verschiedenen Komponenten und Teilprozessen, die beim Lesen eine Rolle spielen, sowie zu deren komplexem Zusammenspiel existieren zahlreiche Modelle, die jeweils unterschiedlichen Komponenten mehr oder weniger viel Gewicht einräumen. Die Theorie der verbalen Effizienz von Perfetti (1989) geht z. B. davon aus, dass das Ausmaß an Leseverständnis in erster Linie von der Sicherheit und Geschwindigkeit der Worterkennung abhängt, während die Theorie des Simple View of Reading von Gough und Tunmer (1986) annimmt, dass das Leseverständnis darüber hinaus bzw. sogar noch vielmehr auch vom Hörverständnis bzw. dem generellen rezeptiven Sprachverständnis beeinflusst wird. Interaktionistische Ansätze wiederum betonen das Zusammenwirken und die wechselseitige Beeinflussung basaler und höherer Leseprozesse (vgl. Christmann & Groeben, 1999; Richter, Christmann, Hurrelmann & Wilkending, 2002). Es wird hier nicht auf alle Theorien im Detail eingegangen. Vielmehr werden die einzelnen Teilprozesse grob differenziert in einerseits die Komponente der basalen Lesekompetenz (Kap. 4.1), deren Prozesse die technische Voraussetzung für Leseverständnis darstellen, und andererseits die Komponente des Textverständnisses (Kap. 4.2), die sich auf die inhaltliche Verarbeitung des Gelesenen bezieht. Anschließend wird die Dimensionalität von Leseverständnis betrachtet und schließlich auf unterschiedliche theoretische Positionen zum Zusammenhang von Lese- und Hörverständnis eingegangen.

4.1 Basale Lesekompetenz

Basale Lesekompetenz ist gegeben, wenn eine Person über die grundlegenden Lesefertigkeiten verfügt, d. h. Wörter und Texte mit einer gewissen Genauigkeit und Geschwindigkeit und ohne große Anstrengung in eine phonologische Repräsentation übersetzen und die entsprechenden Wortbedeutungen abrufen kann (vgl. Auer, Gruber, Mayringer & Wimmer, 2005; D. H. Rost & Buch, 2010). Im Folgenden werden die zentralen Prozesse der basalen Lesekompetenz beschrieben. Dabei wird zunächst auf das Rekodieren und anschließend auf das Dekodieren eingegangen und schließlich die Leseflüssigkeit betrachtet.

Rekodieren. Als Rekodieren bezeichnet man die auf der gelernten Graphem-Phonem-Korrespondenz basierende Übersetzung von Schriftsymbolen auf Buchstaben-

und Wortebene in eine phonologische Repräsentation (H. Marx & Reinhold, 2010). Dafür sind Buchstabenkenntnisse, phonologische Fertigkeiten und Abtaststrategien für Symbolfolgen ebenso notwendig wie Gedächtniszugangs- und -abrufstrategien für Graphem-Phonem-Korrespondenzen. Da im Deutschen die Graphem-Phonem-Korrespondenz weitgehend eindeutig ist, werden bei ihrer Kenntnis die meisten Wörter korrekt gelesen (vgl. Barth, 1999, S. 24).

Beim stillen Lesen bedeutet Rekodieren die Generierung einer internen phonologischen Repräsentation auf der Basis der graphischen und orthographischen Information (H. Marx & Reinhold, 2010). Beim lauten Lesen wird die graphische und orthographische Information artikuliert. Beim Rekodieren ist noch kein Verständnis impliziert. Liest jemand ohne Lateinkenntnisse ein lateinisches Wort vor, rekodiert er es.

Dekodieren. Der Begriff Dekodieren bezeichnet den Abruf der Wortbedeutung aus dem semantischen Lexikon. Der Abruf gelingt bei korrekter Rekodierung, sofern ein entsprechender Eintrag im semantischen Lexikon existiert (H. Marx & Reinhold, 2010). Liest jemand ein lateinisches Wort und kennt die Wortbedeutung, dekodiert er es.

Der Dekodiervorgang erfolgt bei Leseanfängern stets bewusst über den Weg der phonologischen Rekodierung der visuellen Information. Bei geübten Lesern läuft er dagegen meist automatisiert und unbewusst ab. Es wird dabei angenommen, dass geübte Leser Wörter nicht über die serielle Verarbeitung einzelner konkreter Buchstaben erkennen, sondern über die parallele Verarbeitung abstrakter Buchstabeneinheiten (Rayner & Pollatsek, 1989). Zudem wird ein sogenannter „Sichtwortschatz“ aufgebaut, der aus bereits häufig gelesen Wörtern besteht, die mit einem Blick erkannt werden. Analysen der Blickbewegungen geübter Leser zeigen, dass ein Wort fixiert wird und dann ein Sakkadensprung zum nächsten Wort erfolgt. Das Lesen erfolgt also Wort für Wort und nicht mehr Buchstabe für Buchstabe (Klicpera & Gasteiger-Klicpera, 1995, S. 13), was auch als visuelle Worterkennung bezeichnet wird.

Sogenannte „Zwei-Wege-Modelle“ nehmen entsprechend an, dass beim geübten Leser für unvertraute Wörter und Wörter mit regulärer Graphem-Phonem-Zuordnung der phonologische Weg und für vertraute Wörter und Wörter mit irregulärer Aussprache der visuelle Weg gewählt wird (Coltheart, 1978; M. S. Seidenberg, Waters, Barnes & Tanenhaus, 1984). Diese Modelle können jedoch nicht alle beobachteten Phänomene erklären. Es gibt z. B. Hinweise darauf, dass beim lexikalischen Zugriff immer phonologische Komponenten aktiviert werden (z. B. Shankweiler, 1999). Ob und inwieweit die graphemische Information zunächst in eine phonologischen Repräsentation übersetzt werden muss, ist noch unklar (Artelt et al., 2007).

Leseflüssigkeit. Während W. Lenhard (2013, S. 107) unter Leseflüssigkeit lediglich die Lesegeschwindigkeit und -genauigkeit versteht, wird sie von anderen Autoren (z. B. Nix, 2011, S. 55) als Zwischenstufe zwischen basalen Leseprozessen und dem umfassenden Textverstehen eingeordnet. Walter (2013, S. 15) definiert Leseflüssigkeit

als „die Fähigkeit, einen zusammenhängenden Text schnell, reibungslos, prosodisch angemessen, ohne große Anstrengung und automatisiert mit wenig Aufmerksamkeit auf Rekodierprozesse vortragen zu können“. Sie umfasst also Teilfähigkeiten auf Wortebene (Dekodieren und Automatisierung) sowie auf Satz- bzw. lokaler Textebene (Lesegeschwindigkeit und prosodisch phrasiertes Lesen). Somit schließt sie über die basalen Leseprozesse des Rekodierens und Dekodierens hinaus auch Satzverständnis und die Bildung von Zusammenhängen über mehrere aufeinanderfolgende Sätze hinweg sowie beim lauten Lesen eine sinnangemessene Intonation ein (vgl. Nix, 2011, S. 56, 61). Die Leseflüssigkeit wird entsprechend auch als „Brücke“ zwischen basalen Leseprozessen und höheren Leseverständnisleistungen bezeichnet (Nix, 2011, S. 220).

4.2 Textverständnis

Im Folgenden wird das Textverständnis genauer betrachtet, wobei zunächst verschiedene Prozesse der Textverarbeitung beschrieben werden, bevor auf unterschiedliche Repräsentationsebenen eingegangen wird.

4.2.1 Prozesse der Textverarbeitung

Die Prozesse, die bei der inhaltlichen Verarbeitung von Texten ablaufen, lassen sich im Hinblick auf verschiedene Aspekte gliedern. So differenziert man in Bezug auf die Text-Leser-Interaktion zwischen Bottom-up- und Top-down-Verarbeitung, in Bezug auf die Herstellung von Sinnzusammenhängen zwischen lokaler und globaler Kohärenzbildung, in Bezug auf die kognitiven Anforderungen, die die Verarbeitungsprozesse an den Leser stellen, zwischen hierarchieniedrigen und hierarchiehöheren Prozessen sowie zwischen verschiedenen Formen von Inferenzen.

Text-Leser-Interaktion. In der Forschung ist man sich weitgehend einig, dass es sich beim Textverstehen um einen aktiven, kognitiv-konstruktivistischen Prozess handelt, bei dem Leser und Text miteinander interagieren und der zum Aufbau einer mentalen Repräsentation semantischer Textstrukturen führt (vgl. z. B. Adam-Schwebe et al., 2009; D. H. Rost & Buch, 2010). Bedeutungsinhalte werden nicht nur durch die Rekonstruktion des semantischen Gehalts einzelner Wörter und deren Verknüpfung mit den sie umgebenden Kontextinformationen rekonstruiert, sondern auch durch die Interpretation auf der Grundlage persönlichen Vorwissens (Christmann & Groeben, 1999). W. Kintsch (1998, S. 93ff.) beschreibt den Leseprozess als eine Kombination aus textgesteuertem Konstruktions- und vorwissensgesteuertem Integrationsprozess. Die beiden Verarbeitungsprozesse stehen in einem Interdependenzverhältnis zueinander und werden häufig als Bottom-up- und Top-down-Prozesse bezeichnet (D. H. Rost & Buch, 2010). Erst durch ihr Zusammenwirken wird Verständnis möglich.

Bei *Bottom-up-Prozessen* geht die Verarbeitung vom Reiz aus und unterliegt kaum willentlicher Kontrolle. Sie läuft meist implizit, automatisiert und kontextunabhängig ab (D. H. Rost & Buch, 2010). Die sukzessive Verarbeitung visueller Reize führt schließlich zur kognitiven Repräsentation eines Textes. Vom Verstehen auf Wortebene über die Satzebene wird einem Text Schritt für Schritt Sinn entnommen. Bottom-up-Prozesse kommen vor allem dann zum Tragen, wenn ein Leser kein oder wenig Vorwissen zu einem Thema hat oder sein Vorwissen nicht nutzen kann. Sie setzen lediglich lexikalische und linguistische Kenntnisse voraus. Der zweite Verarbeitungsweg, die *Top-down-Prozesse*, erfordern hingegen ein gewisses Maß an Vorwissen, das aus thematischem Vorwissen, bereits Gelesenem, allgemeinem Weltwissen sowie der Syntax und der Textstruktur bestehen und zur Steuerung untergeordneter Prozesse sowie der Aufmerksamkeit beitragen kann (vgl. D. H. Rost & Buch, 2010). Neu „Erlesenes“ wird in vorhandene Wissensstrukturen integriert.

Kohärenzbildung. Unter Kohärenzbildung versteht man das Herstellen von Sinnzusammenhängen über bestimmte Textabschnitte hinweg. Man differenziert dabei zwischen lokaler Kohärenzbildung auf Satz- und satzübergreifender Ebene und globaler Kohärenzbildung auf Textebene.

Zum Verstehen eines Satzes genügt es nicht, die einzelnen Wörter zu verstehen. Es muss darüber hinaus die syntaktische Struktur erfasst und der semantische Gehalt der einzelnen Wörter in einen Zusammenhang gebracht werden (W. Lenhard & Artelt, 2009). Auf der Basis ihrer semantischen Relationen werden Wortsequenzen zu Sinneinheiten integriert und es entsteht eine propositionale Repräsentation des Satzes (van Dijk & Kintsch, 1983; W. Kintsch, 1998). Für das Verständnis eines Textabschnittes ist es weiter nötig, den Inhalt mehrerer aufeinanderfolgender Sätze durch die Analyse ihrer semantischen Relationen miteinander zu verknüpfen (W. Lenhard & Artelt, 2009; Richter et al., 2002). Leser greifen bei der Verknüpfung der Propositionen meist automatisch auf ihr inhaltsbezogenes oder allgemeines Weltwissen zurück (Artelt et al., 2007). Häufig unterstützen zudem Bindeglieder und Verweise zwischen Sätzen, sogenannte „sprachliche Kohäsionsmittel“, diesen Vorgang (Christmann & Groeben, 1999; W. Lenhard & Artelt, 2009). Kohäsionsmittel können beispielsweise Rückverweise, Vorverweise, Wortwiederholungen oder Wiederaufnahmen von Satzsequenzen durch sogenannte „Pro-Formen“ („dies“, „das“, „so“) sein. All diese Prozesse dienen der *lokalen Kohärenzbildung*.

Um einen ganzen Text verstehen zu können, müssen darüber hinaus *globale Kohärenzen* gebildet werden, d. h. größere Textabschnitte müssen analysiert, als Folge von inhaltlichen Aussagen verstanden, verdichtet und miteinander verknüpft werden (W. Lenhard & Artelt, 2009). Graesser, Singer und Trabasso (1994) gehen diesbezüglich davon aus, dass ein Leser zur Herstellung globaler Kohärenz in seinem Langzeitgedächtnis nach passenden Informationen sucht und diese ins Arbeitsgedächtnis zieht, um sie dort mit den Textinformationen sinnvoll zu integrieren. Dadurch ent-

stehen Makropropositionen, die sich schließlich zu Makrostrukturen zusammenfügen (van Dijk, 1980, S. 41).

Kognitive Anforderung. In Abhängigkeit von den kognitiven Anforderungen wird beim Leseverständnis zwischen hierarchieniedrigen und hierarchiehohen Prozessen differenziert. Als *hierarchieniedrig* zählen der Aufbau einer Textrepräsentation auf Basis der Worterkennung, der Wortsequenzen sowie lokaler Kohärenzen (Artelt et al., 2007). Bei geübten Lesern laufen diese Prozesse automatisch ab und beanspruchen daher wenig kognitive Kapazität.

Als *hierarchiehoch* gilt demgegenüber die Bildung von globaler Kohärenz, von Inferenzen und von Schemata zur globalen Textordnung sowie das Erkennen rhetorischer Strategien (Artelt et al., 2007; W. Lenhard & Artelt, 2009). Diese Prozesse sind kognitiv anspruchsvoller, laufen weniger automatisiert ab und erfordern bewusste geistige Anstrengung (W. Lenhard & Artelt, 2009; Rosebrock & Nix, 2011). Durch die Anwendung von Lesestrategien ist eine zweckorientierte Steuerung hierarchiehoher Prozesse möglich (W. Lenhard & Artelt, 2009). Daher spielen metakognitive Fähigkeiten, das schemageleitete Textverstehen und das Vorwissen hierbei eine größere Rolle. Der Text wird sowohl reduktiv als auch elaborativ verarbeitet, d. h. im Text enthaltene Informationen werden z. B. selektiert, generalisiert oder weiterführendes Denken initiiert. Mithilfe dieser Prozesse kommt es zu einem abstrakteren Verständnis des globalen Sinnzusammenhangs eines Textes (W. Lenhard & Artelt, 2009; Rosebrock & Nix, 2011).

Inferenzen. Um Texte verstehen und Kohärenzen bilden zu können, ist es nötig, Inferenzen zu generieren und neue Informationen aus dem Text mit vorhandenem Wissen zu verknüpfen. Dabei werden verschiedene Formen von Inferenzen unterschieden (Überblick s. Graesser et al., 2007), z. B. referenzielle Inferenzen, die beispielsweise ein Pronomen oder eine Nominalphrase mit einem bereits genannten Referenten in Beziehung setzen, oder rückwärtige Inferenzen, die den aktuell gelesenen Textabschnitt mit vorherigen Abschnitten verknüpfen (Artelt et al., 2007; Rizzella & O'Brien, 1996). Weiter gibt es prädiktive Inferenzen der Antizipation und Hypothesenbildung bezüglich weiterer Textinhalte, die auf dem bisher Gelesenen basieren und die Integration neuer Informationen in die aktuelle Textrepräsentation erleichtern (Unsöld, 2008, S. 67). Elaborative Inferenzen reichern die Textrepräsentation mit Informationen aus dem Langzeitgedächtnis an (Artelt et al., 2007; van den Broek, Fletcher & Risden, 1993). Kausale Inferenzen spielen in Lehr- und Sachtexten sowie Erzählungen eine wichtige Rolle und sind schon für das Verständnis einfacher Satzsequenzen zur Rekonstruktion impliziter Prämissen notwendig (Artelt et al., 2007; Trabasso & van den Broek, 1985). Das Verstehen der Satzsequenz „Susanne legte das Eis in die Sonne. Das Eis schmolz." erfordert z. B. die Rekonstruktion der Prämisse, dass Eis in der Sonne schmilzt (vgl. Artelt et al., 2007; Singer, Halldorson, Lear & Andrusiak, 1992).

Graesser et al. (1994) entwickelten ein Modell der Inferenzbildung, das ein „Prinzip der Suche nach Bedeutung“ postuliert. Es geht von einem prinzipiellen Ziel des Lesers aus, einem Text Bedeutung zu entnehmen. Im Einzelnen wird angenommen, dass ein Leser versucht, (1) seine Leseziele zu befriedigen, (2) lokale und globale Kohärenz herzustellen sowie (3) im Text dargestellte Handlungen und Ereignisse zu erklären. Die erste Annahme bezieht sich darauf, dass Leser mit dem Lesen unterschiedliche Ziele verfolgen können. Sie können z. B. zur Unterhaltung lesen oder um Informationen zu gewinnen. Abhängig vom Ziel wird der Leser eine andere Herangehensweise wählen. Er wird die Verarbeitungstiefe anpassen, zielorientiert Informationen entnehmen und Inferenzen bilden. Auch Faktoren wie die für die Verarbeitung zur Verfügung stehende Zeit oder angekündigte Verständnisfragen beeinflussen die Verarbeitungstiefe und Inferenzbildung (vgl. Diergarten, 2010, S. 79). Die zweite Annahme (Kohärenzbildung) bezieht sich auf das Streben des Lesers nach der Generierung einer umfassenden Bedeutungsrepräsentation, die sowohl Informationen aus dem Text als auch das vorhandene Weltwissen impliziert. Diese Generierung kann textseitig durch entsprechende Gestaltung unterstützt werden, leserseitig sind Vorwissen und klare Leseziele nötig (vgl. Diergarten, 2010, S. 79f.). Die dritte Annahme – dass der Leser versucht, sich die Handlungen und Ereignisse im Text zu erklären – postuliert, dass „Warum“-Fragen den Leser bei der Inferenzbildung leiten. Der Leser fragt sich demnach, warum bestimmte Ereignisse im Text dargestellt werden und warum der Autor bestimmte Informationen gibt. Für ungeplante Ereignisse bieten sich dabei z. B. kausale Ursachen und vorangegangene Zustände als Erklärung an, für willentliche Handlungen die Ziele des Protagonisten (Diergarten, 2010, S. 80).

4.2.2 Ebenen der Textrepräsentation

Abhängig von der Verarbeitungstiefe werden außerdem verschiedene Ebenen mentaler Textrepräsentationen unterschieden (z. B. W. Kintsch, 1998; Zwaan & Kaschak, 2009; Zwaan & Radvansky, 1998; Zwaan, Radvansky, Hilliard & Curiel, 1998; Schnotz & Dutke, 2004; Graesser, Millis & Zwaan, 1997). Die meisten Theorien zur Bildung mentaler Textrepräsentationen stützen sich auf die sehr ähnlichen Modelle von van Dijk und Kintsch (1983) sowie Johnson-Laird (1983). Diese unterscheiden jeweils drei Ebenen der Textrepräsentation: (1) Eine Repräsentation der Textoberfläche, die den genauen Wortlaut enthält, (2) eine Repräsentation des Textinhaltes, die auch als Textbasis oder propositionale Repräsentation bezeichnet wird, und (3) eine Repräsentation, die über den expliziten Textinhalt hinausgeht und von van Dijk und Kintsch (1983, S. 12) als „Situationsmodell“ („situation model“) und von Johnson-Laird (1983, S. 10ff.) als „mentales Modell“ („mental model“) bezeichnet wird. Diese drei Ebenen werden im Folgenden genauer beschrieben, ebenso wie die sogenannte „Fuzzy Trace-Theorie“ (Brainerd & Reyna, 1991) und zwei Erweiterungen um Metaebenen der Textrepräsentation.

(1) Textoberfläche. Die Repräsentation der Textoberfläche eines Satzes enthält Informationen über graphemische, lexikalische und syntaktische Texteigenschaften, also den exakten Wortlaut, die semantische Bedeutung der einzelnen Wörter und die syntaktische Struktur des Satzes (vgl. z. B. Schnotz & Dutke, 2004). Ihr Aufbau läuft bei geübten Lesern automatisch ab. Mit Rekognitionsaufgaben kann geprüft werden, ob eine Repräsentation der Textoberfläche aufgebaut wurde. Dabei können z. B. verschiedene Sätze gezeigt werden, die im Wortlaut (Oberflächenstruktur), aber nicht in der Bedeutung (Semantik) variieren. So zeigte Sachs (1967), dass Probanden Abweichungen des Testsatzes vom Originalsatz seltener bemerkten, wenn sich die Sätze nur in ihrer Oberflächenstruktur unterschieden, als wenn auch ein semantischer Unterschied gegeben war. Allerdings war dieser Effekt nur zu beobachten, wenn zwischen dem Originalsatz und dem Testsatz mindestens 80 Silben lagen. Lagen weniger Silben dazwischen, wurden Abweichungen in der Oberflächenstruktur und in der semantischen Struktur gleich häufig entdeckt. Diese Beobachtung wurde als Hinweis darauf interpretiert, dass Informationen auf der Textoberfläche nur flüchtig gespeichert werden (Bransford, Barclay & Franks, 1972).

Spätere Studien zeigten jedoch, dass auch im Langzeitgedächtnis noch Informationen zur Textoberfläche gespeichert sind. So ließen Tardif und Craik (1989) Probanden verschiedene Texte zum Teil in der Originalversion und zum Teil in einer paraphrasierten Version laut vorlesen. Bei einer Wiedervorlage der Texte eine Woche später wurden die Versionen teilweise vertauscht, was dazu führte, dass die Probanden in diesen Fällen länger zum Lesen brauchten. Der Befund spricht dafür, dass die Oberflächenstruktur des Textes zumindest teilweise noch repräsentiert war.

(2) Textbasis. Auf der Ebene der Textbasis wird der vom Wortlaut unabhängige semantische Gehalt des Textes in Form von abstrakten Bedeutungseinheiten, sogenannten „Propositionen“, gespeichert. Da sich die propositionale Ebene ebenso wie die Textoberfläche ausschließlich auf explizit im Text enthaltene Informationen bezieht, werden diese beiden Repräsentationsformen auch als „textbasiert“ bezeichnet.

Jede Proposition besteht aus einem Prädikat und Argumenten, wobei das Prädikat die Argumente zueinander in Relation setzt (Christmann, 2004). An der Textoberfläche treten Prädikate als Verben, Adjektive, Präpositionen oder Adverbien auf und beschreiben z. B. Zustände, Ereignisse oder Eigenschaften. Argumente sind in der Regel durch Nomen vertreten und legen fest, worüber (z. B. Objekte, Personen oder Sachverhalte) etwas ausgesagt wird. Dies soll anhand folgenden Beispielsatzes nach Fletcher (1994, S. 589, eigene Übersetzung) verdeutlicht werden: „Der Frosch fraß die Fliege.“ Hier ist „Frosch“ der Agent und „Fliege“ das Objekt. Beide stellen die Argumente des Satzes dar. „Fressen“ ist das Prädikat, das die Argumente miteinander verknüpft.

Aus einer propositionalen Repräsentation geht dabei z. B. nicht hervor, ob der Satz im Aktiv oder im Passiv formuliert war. Auch durch Pronomen und Synonyme würde

sich die propositionale Struktur des Satzes nicht verändern. Nicht alle Propositionen sind jedoch so simpel wie das Beispiel. Temporale und/oder lokale Relationen können die Komplexität von Propositionen erhöhen (vgl. W. Kintsch, 1998, S. 37ff.).

(3) Situationsmodell. Das Situationsmodell bezieht sich auf die im Text dargestellte Situation und geht im Gegensatz zu den beiden bisher beschriebenen Ebenen über die explizit im Text enthaltenen Informationen hinaus. Im Sinne eines konstruktiven Prozesses werden Verknüpfungen der propositionalen Struktur mit Vorwissen aus dem Langzeitgedächtnis hergestellt. Nicht im Text enthaltene Informationen werden aus dem Kontext und unter Zuhilfenahme dieses Vorwissens erschlossen (Johnson-Laird, 1983; W. Kintsch, 1998). Auf diese Weise entsteht im Kopf des Lesers ein umfassendes Bild der im Text beschriebenen Situation, das vom individuellen Weltwissen des Lesers beeinflusst ist. Das Situationsmodell kann somit als „analoge, inhaltsspezifische und anschauliche mentale Repräsentation des im Text dargestellten Sachverhalts“ beschrieben werden (W. Lenhard & Artelt, 2009, S. 10).

In Bezug auf den Beispielsatz von Fletcher (1994) werden beim Leser unterschiedliche Assoziationen ausgebildet. Jeder Leser sieht vor seinem inneren Auge z. B. einen etwas anderen Frosch, der bei dem einen am See, bei dem anderen am Gartenteich sitzt. Abhängig vom Vorwissen des Lesers kann die beschriebene Situation in der Vorstellung des Lesers durch nicht explizit im Text enthaltene Details angereichert werden.

Da die genaue Struktur der Sprachäußerung im Situationsmodell nicht gespeichert wird, fällt bei einer Abfrage die Wiedergabe weniger genau aus, es handelt sich eher um eine freie Beschreibung (Schnotz & Dutke, 2004). Van den Broek und Gustafson (1999) sehen die Textbasis und das Situationsmodell als zwei Pole eines Kontinuums. Abhängig vom Ausmaß des vorhandenen Vorwissens wird eine stärker textbasierte oder eine freier interpretierte Repräsentation generiert.

Nach Zwaan et al. (1998) lassen sich bei Situationsmodellen fünf Dimensionen differenzieren: Raum, Zeit, Protagonist, Kausalität und Intentionalität. Zur Untersuchung der Dimensionen wurde ein sogenanntes „multidimensionales Event-Indexing-Modell“ entwickelt (Zwaan, Magliano & Graesser, 1995). Dieses postuliert, dass beim Lesen eines Textes beim ersten zu verarbeitenden Ereignis für alle fünf Dimensionen jeweils ein Index gebildet wird. Alle Indizes werden bei jedem weiteren Ereignis auf ihre Haltbarkeit geprüft und nötigenfalls angepasst. Ändert sich in einem neuen Satz z. B. das Ziel (Intentionalität), muss der Zielindex angepasst werden; tritt ein neuer Protagonist hinzu, muss der Protagonistenindex entsprechend geändert werden; etc. Es zeigte sich, dass mit der Anzahl der zu aktualisierenden Dimensionen die Lesezeit für einen Satz anstieg, was als Folge des Verbrauchs kognitiver Ressourcen für die Aktualisierung der Indizes interpretiert wurde. Alle fünf Dimensionen wurden bereits empirisch bestätigt (Zwaan et al., 1995, 1998).

Fuzzy Trace-Theorie. Obwohl sich zeigte, dass bereits jüngere Kinder über alle drei Repräsentationsebenen verfügen, scheinen Kinder im Vor- und Grundschulalter Informationen überwiegend auf Ebene der Textoberfläche zu repräsentieren, während zwischen sieben und elf Jahren ein Übergang zur Präferenz der Speicherung in Form eines Situationsmodells stattfindet (Nieding, 2006; Ganea, Shutts, Spelke & DeLoache, 2007). Als Erklärung für den Übergang dient die sogenannte „Fuzzy Trace-Theorie" von Brainerd und Reyna (1998, 2002; Reyna & Brainerd, 1995). Das Modell differenziert zwischen einem Gedächtnis für das Wesentliche und einem Gedächtnis für das Wörtliche. Es wird ein Kontinuum vom Wesentlichen zum Wörtlichen angenommen, wobei das Wesentliche in Form von „Fuzzy Traces" gespeichert wird. Das wörtliche Erinnern von Informationen hat für Kinder im Spracherwerbsalter einen adaptiven Vorteil, da es ihnen ermöglicht, täglich bis zu zehn neue Wörter zu lernen (Gleitman & Gleitman, 1992). Ab etwa einem Alter von sieben Jahren findet der Wechsel hin zur Präferenz für die Speicherung in Form von Fuzzy Traces statt, die im Vergleich zur wörtlichen Repräsentation stabiler und leichter abrufbar sind.

Metaebenen. Nach Graesser et al. (1997), Schnotz und Dutke (2004) sowie Richter et al. (2002) existieren über die drei genannten Repräsentationsebenen (Textoberfläche, Textbasis, Situationsmodell) hinaus zwei weitere Ebenen des Textverständnisses: (4) die Ebene des Textgenres bzw. der Superstrukturen und (5) die Ebene der Darstellungsstrategien bzw. der Kommunikation oder rhetorischer Strategien. Diese sind beide auf der Metaebene angesiedelt und beziehen sich nicht direkt auf die inhaltlichen Informationen des Textes, sondern auf formale Aspekte und sprachliche Mittel, die zur Inferenzbildung und zum Schließen von Verständnislücken herangezogen werden und eine angemessene Textinterpretation ermöglichen.

(4) Bei der Verarbeitung auf *Ebene des Textgenres* geht es um den Zusammenhang zwischen strukturellen Textmerkmalen, dem vom Autor gewählten Textgenre und bestimmten allgemeinen Kommunikationsformen, z. B. des Erzählens, Berichtens, Erklärens etc. (Schnotz & Dutke, 2004). Richter et al. (2002) bezeichnen die Konstruktion einer inhaltsunabhängigen, globalen Ordnung von Texten, die als Regeln oder Kategorien gespeichert werden, als die Bildung von sogenannten „Superstrukturen". Die verschiedenen Textsorten stellen unterschiedliche Anforderungen an den Leser und führen gemäß der Superstrukturen zu genrespezifischen Erwartungen des Lesers an den Text sowie zur Bildung von Hypothesen (Graesser et al., 1997; Rosebrock & Nix, 2011).

(5) Die *Ebene der Darstellungsstrategien* bezieht sich auf den pragmatischen kommunikativen Kontext sowie das Erkennen rhetorischer, stilistischer und argumentativer Strategien und hängt mit der Intention des Autors zusammen bzw. bringt diese zum Ausdruck (Richter et al., 2002; Rosebrock & Nix, 2011).

Der Leseprozess wird also bei geübten Lesern auch aus der Metaperspektive begleitet. Die Prozesse auf den Metaebenen werden als hierarchiehöchste kognitive Ver-

arbeitungsprozesse betrachtet. Zur Verdeutlichung der Metaebenen wird häufig auf Märchenparodien verwiesen, die voraussetzen, dass der Leser mit konventionellen Strukturen von Märchen vertraut ist, um dann die dadurch entstehenden Erwartungen zu durchkreuzen (Rosebrock & Nix, 2011, S. 15). Ein Leser, der diese Strategie des Autors erkennt, wird durch sie nicht verwirrt, sondern genießt das Spiel mit Erzählkonventionen.

4.3 Dimensionalität von Leseverständnis

Die Darstellung der Vielzahl der am Lesen beteiligten Prozesse in den vergangenen Abschnitten führt nun zur Frage, aus wie vielen distinkten Dimensionen sich das Konstrukt Leseverständnis zusammensetzt. Hier existieren sowohl Befunde, die für Zwei- oder sogar Dreidimensionalität sprechen, als auch Befunde, die Eindimensionalität nahelegen.

Simple View of Reading. Der von Gough et al. (Gough & Tunmer, 1986; Hoover & Gough, 1990; Gough et al., 1996) beschriebene Ansatz des „Simple View of Reading“ (SVR) postuliert, dass das Leseverständnis von den zwei Faktoren (Dimensionen) basale Lesefertigkeit und Hörverstehen abhängt, wobei beide gleich zu gewichten sind. Der SVR greift damit die Vorstellung von Stanovich et al. (Stanovich & West, 1989; Stanovich, 1991; Stanovich & Siegel, 1994) auf, dass Hörverstehen ein bedeutsamerer Prädiktor für das Leseverständnis ist als die Intelligenz.

Unter basaler Lesefertigkeit werden beim SVR phonologisches Rekodieren und effizientes Dekodieren gefasst (H. Marx & Jungmann, 2000). Laut Gough und Tunmer (1986) ist zu Beginn des Schriftspracherwerbs hauptsächlich das Rekodieren von Bedeutung, welches im weiteren Verlauf jedoch immer stärker von effizienteren, nicht näher spezifizierten Dekodierfähigkeiten abgelöst wird (Hoover & Gough, 1990). Mit Hörverstehen ist die Fähigkeit gemeint, die syntaktische Struktur und Bedeutung eines über das Ohr aufgenommenen Satzes zu erfassen und mit den vorangehenden und nachfolgenden Sätzen zu verknüpfen. Es wird angenommen, dass die Prozesse des Verstehens bei der Verarbeitung schriftlicher und auditiver Texte im Wesentlichen identisch sind, trotz Unterschieden z. B. im Hinblick auf die Sprache (formell vs. informell) oder die Darbietungsmodalität (Interpunktion bei schriftlichem, Intonation bei auditivem Material).

Unterschiede in Verständnisleistungen zwischen den Darbietungsmodalitäten werden dabei auf modalitätsspezifische Faktoren zurückgeführt. Im Falle des Lesens wäre das z. B. die Fähigkeit, Buchstabensequenzen zu dekodieren. Bei maximaler Ausprägung des Hörverstehens wird das Leseverstehen allein von der Fähigkeit bestimmt, einzelne Wörter zu entschlüsseln (Hoover & Gough, 1990). Das Leseverstehen resultiert demnach aus einer multiplikativen Verknüpfung von Hörverstehen und Deko-

dierfähigkeit, da beide Fähigkeiten vorhanden sein müssen, um die Bedeutung eines Textes zu erfassen.

Dem SVR zufolge kann defizitäres Leseverständnis entweder auf Probleme beim Hörverständnis zurückgeführt werden, die auf grundlegenden Sprachleistungsdefiziten basieren (z. B. geringer Wortschatz, niedrige verbale Intelligenz oder auch reduzierte allgemeine kognitive Leistungsfähigkeit) oder auf Probleme bei lesespezifischen Faktoren (z. B. eine unzureichende Dekodierfähigkeit oder durch Übung nur in geringem Maße beeinflussbare Geschwindigkeitsfaktoren). Darüber hinaus können Probleme beim Leseverständnis durch mangelnde Aufmerksamkeit für den Verständnisprozess hervorgerufen werden, da die Dekodierung bereits einen wesentlichen Teil der Aufmerksamkeit in Anspruch nimmt. Findet das Dekodieren automatisiert statt, ist zu erwarten, dass Lese- und Hörverständnis ähnlich stark ausgeprägt sind. Demzufolge könnte das Hörverständnis als Indikator für die potenziell erreichbare Leistung beim Leseverständnis dienen (Carlisle & Felbinger, 1991; D. H. Rost & Buch, 2010). Hoover und Gough (1990) unterscheiden daher gemäß der Ausprägung (hoch vs. niedrig) auf den beiden Dimensionen Dekodierleistung und Hörverständnis vier Subgruppen von Lesern, wobei drei Gruppen von schwachen Lesern gebildet werden (s. Abb. 2). Sogenannte „Garden-Variety-Poor-Reader“ verfügen demnach weder über eine hohe Dekodierfähigkeit noch über ein hohes Hörverständnis. Entsprechend sind diese Personen generell leseschwach. Hyperlexie liegt vor, wenn eine Person zwar über eine gute Dekodierfähigkeit verfügt, jedoch den Sinn des Gelesenen nicht erfassen kann. Bei Dyslexie weisen Personen trotz hohen Hörverstehens aufgrund mangelhafter Dekodierfähigkeit nur ein schwaches Leseverständnis auf. Ist sowohl das Hörverstehen als auch die Dekodierfähigkeit hoch ausgeprägt, verfügt eine Person über ein gutes Leseverständnis.

		Dekodierleistung	
		niedrig	hoch
Hörverstehen	niedrig	Garden-Variety-Poor-Reader	Hyperlexie
	hoch	Dyslexie	Gut ausgeprägtes Leseverständnis

Abbildung 2. Subgruppen von Lesern nach dem SVR.

Sowohl für die isolierte Schwäche im Leseverständnis als auch für die isolierte Schwäche der Dekodierleistung gibt es inzwischen empirische Befunde, wobei letztere seltener auftritt (vgl. Catts, Adlof & Weismer, 2006; Stothard & Hulme, 1995;

Ennemoser et al., 2012). Catts et al. (2006) zeigten, dass bei der Gruppe der schwachen Dekodierer eher phonologische Defizite vorliegen, während bei der Gruppe der Personen mit schwacher Leseverständnisleistung eher das generelle Sprachverständnis beeinträchtigt ist.

Eine Metaanalyse von Florit und Cain (2011) ergab, dass das Sprachverständnis vor allem bei Lesern transparenter Orthographien ein wichtiger Prädiktor für das Leseverständnis war und bei diesen sogar bei Leseanfängern stärkeren Einfluss auf das Leseverständnis hatte als das Dekodieren. Dabei ist jedoch zu berücksichtigen, dass dies nur auf die Dekodier*genauigkeit* zutrifft. Sprachverständnis war kein besserer Prädiktor als Dekodier*flüssigkeit*. So zeigte beispielsweise Wimmer (1993) bei deutschen Kindern mit Dyslexie, dass schlechte Dekodierfähigkeiten vor allem in der Dekodierflüssigkeit und weniger in der Dekodiergenauigkeit bestehen.

Inzwischen wurde das SVR-Modell von einigen Forschern modifiziert und erweitert. Carver (1993) schlug mit dem Modell „Simple View of Reading II“ vor, die Lesegeschwindigkeit zusätzlich aufzunehmen. Auch Aaron, Joshi und Williams (1999), Adlof, Catts und Little (2006) sowie Spear-Swerling (2006) fügten eine Flüssigkeitskomponente hinzu. Joshi und Aaron (2000) konnten in einer Untersuchung durch die Hinzunahme der Dekodiergeschwindigkeit bei der Erklärung von Leseverständnis die Varianzaufklärung um 10 % von 47,6 % auf 57,8 % steigern, und Kershaw und Schatschneider (2012) fanden in ihrer Studie ebenfalls eine zusätzliche Bedeutung der Leseflüssigkeit. Georgiou, Das und Hayward (2009), Johnston und Kirby (2006) sowie Tiu, Thompson und Lewis (2003) nahmen die Verarbeitungsgeschwindigkeit in das Modell mit auf. Cain, Oakhill und Bryant (2004), Seigneuric und Ehrlich (2005) sowie Seigneuric, Ehrlich, Jane und Yuill (2000) fügten die Arbeitsgedächtnisleistung hinzu. Insgesamt hat sich also vor allem eine – wenn auch unterschiedlich benannte (Lesegeschwindigkeit, Leseflüssigkeit, Dekodiergeschwindigkeit, Verarbeitungsgeschwindigkeit) – Geschwindigkeitskomponente mehrfach als wichtiger zusätzlicher Prädiktor für das Leseverständnis sowohl in der Grundschule als auch auf den weiterführenden Schulen erwiesen.

Eindimensionalität von Leseverständnis. Es gibt jedoch auch Untersuchungen, die dem SVR zu widersprechen scheinen (Gough et al., 1996). Dazu zählen faktorenanalytische Ergebnisse zum Leseverstehen, die nur einen einzigen Faktor (eine einzige Dimension), nämlich „allgemeines Verständnis“ und keinen zweiten Faktor „Dekodierleistung“ fanden (z. B. Carroll, 1988; D. H. Rost, 1987, 1989; D. H. Rost, Czeschlik & Van der Kooij, 1986; R. Zwick, 1987; Andrich & Godfrey, 1979).

Hinzu kommt, dass unterschiedliche Skalen des Leseverständnisses nach einer Minderungskorrektur[4] stets sehr hoch miteinander korrelieren (D. H. Rost, 1993). Bei

[4] Korrekturformel zur Berücksichtigung der Messungenauigkeit von Skalen bei Korrelationen

PISA und IGLU[5] konnten beispielsweise die zugrunde gelegten ausdifferenzierten Lesekompetenzmodelle mit mehreren distinkten Skalen nicht empirisch bestätigt werden. Die latenten Interkorrelationen zwischen den verschiedenen Facetten von Lesekompetenz lagen bei IGLU durchgehend über $r = .87$, bei PISA sogar über $r = .90$. Die PISA-Autoren selbst schreiben, dass es sich wohl eher um Subskalen einer einzigen Lesekompetenz handelt als um mehrere distinkte Skalen (Artelt et al., 2001, S. 134). A. Voss (2006) kommt ebenfalls zu dem Schluss, dass sich ein ausdifferenziertes Modell des Leseverständnisses, das mehrere distinkte Dimensionen umfasst trotz großem Aufwand bei der Skalenkonstruktion und trotz versierter Auswertungsmethoden bisher nicht empirisch nachweisen ließ. Laut D. H. Rost und Buch (2010) lassen die Befunde darauf schließen, dass Leseverständnis nichts anderes ist als die Lösung verbal kodierter Problemstellungen, wobei Wortschatz, Intelligenz und Denkfähigkeit die Schlüsselvariablen darstellen. Daher würde es keinen Sinn machen, Leseverständnisprofile auf der Basis verschiedener Subskalen bzw. Teilfertigkeiten zu erstellen oder differenzialdiagnostisch unterschiedliche Typen des Leseverständnisses zu identifizieren.

Zu bedenken ist jedoch, dass viele der Untersuchungen, die nur eine einzige Dimension fanden, klären sollten, ob Lese*verstehen* eine einheitliche Fertigkeit darstellt und entsprechend zu operationalisieren ist (vgl. H. Marx & Jungmann, 2000). Rekodier- und Dekodierleistungen, die dem SVR zufolge die zweite Dimension darstellen, blieben weitgehend außen vor, oder Dekodier- und Verständnisleistungen waren konfundiert. Demnach wurde in diesen Untersuchungen nur die Komponente von Leseverständnis erfasst, die auch gemäß dem SVR eindimensional ist. Die Ergebnisse sind somit eigentlich gar nicht widersprüchlich.

Lese- und Hörverständnis. Sowohl der SVR als auch die Position, die bezüglich des Leseverständnisses von einer einzigen Dimension ausgeht, nehmen an, dass es sich bei den für das Leseverstehen verantwortlichen Verständnisprozessen eher um generelles rezeptives Sprachverständnis oder um verbale Intelligenz als um lesespezifische Prozesse handelt. Demzufolge sollten die gleichen Prozesse wie beim Hörverstehen beteiligt sein. Diese Annahme modalitätsunabhängiger Verstehensprozesse bei der rezeptiven Sprachverarbeitung wird auch als „monistische Position“ bezeichnet. Demgegenüber steht jedoch die sogenannte „dualistische Position“, die davon ausgeht, dass sich die Verstehensprozesse bei der Verarbeitung schriftlichen Materials grundsätzlich von jenen Verstehensprozessen unterscheiden, die für das Hörverständnis relevant sind.

[5] Die *Progress in International Reading Literacy Study* (PIRLS) wird in Deutschland *Internationale Grundschul-Lese-Untersuchung* (IGLU) genannt und wird von der *International Association for the Evaluation of Educational Achievement* (IEA) alle fünf Jahre durchgeführt. Sie erfasst die Lesekompetenz von Schülern am Ende der Grundschulzeit, d. h. am Ende der vierten Klasse bzw. typischerweise bei Schülern im Alter von ca. neun bis zehn Jahren.

Die *monistische Position* geht von modalitätsunabhängigen Verstehensprozessen aus, die nach der Worterkennung für das Lese- und das Hörverständnis identisch sind (vgl. D. H. Rost & Buch, 2010; H. Marx & Jungmann, 2000). Unterschiede in der Verständnisleistung beim Lesen und Hören entstehen demnach aufgrund der lesespezifischen Fertigkeiten (z. B. Dekodierleistung). Ist das Hörverstehen voll ausgeprägt, sollte das Satz- und Textverstehen auch kein Problem sein – Voraussetzung ist lediglich eine ausreichende Re- und Dekodierleistung (vgl. H. Marx & Jungmann, 2000).

Der enge Zusammenhang zwischen Hör- und Leseverstehen wurde inzwischen in zahlreichen Studien nachgewiesen (D. H. Rost & Hartmann, 1992; H. Marx & Jungmann, 2000; Rupley, 1997; Mommers, 1987). Während sich beim Hörverstehen ein kontinuierlicher moderater Anstieg von Klassenstufe zu Klassenstufe zeigt, scheint das Leseverständnis zwischen dem Ende der ersten und der Mitte der zweiten Klasse vergleichsweise stark anzusteigen, was auf die Automatisierung des Dekodierprozesses zurückgeführt wird (H. Marx & Jungmann, 2000). Gerade bei Leseanfängern im deutschsprachigen Raum zeigte sich, dass die Dekodierfähigkeit das Leseverständnis determiniert. Ab der zweiten Klasse ist der Zusammenhang zwischen Hör- und Leseverstehen substanziell und liegt zwischen $r = .60$ und $r = .65$ (H. Marx & Jungmann, 2000; D. H. Rost & Hartmann, 1992). Bis zur sechsten Klasse konnte der Einfluss der Dekodierleistung gezeigt werden (D. H. Rost & Hartmann, 1992; H. Marx & Jungmann, 2000; Mommers, 1987; Rupley, 1997). Mit steigender Lesefertigkeit bzw. Automatisierung des Leseprozesses scheint der Einfluss lesespezifischer Faktoren abzunehmen und die Diskrepanz zwischen dem Hör- und Leseverstehen zu sinken. Während die Korrelationen im Schulalter generell zwischen $r = .45$ und $r = .75$ liegen, zeigt sich bei Erwachsenen ein deutlich höherer Zusammenhang bzw. kaum noch ein Unterschied (D. H. Rost & Buch, 2010). Insgesamt korreliert die Leistung im Leseverständnis mit $r = .62$ bzw. minderungskorrigiert mit $r = .85$ sehr hoch mit der im Hörverständnis. Diese Befunde deuten auf eine holistische Verarbeitung hin, bei der das sprachlogische Denken eine prominente Rolle spielt (D. H. Rost & Hartmann, 1992; H. Marx & Jungmann, 2000).

Die *dualistische Position* postuliert demgegenüber modalitätsspezifische Faktoren der Informationsverarbeitung. Sie geht davon aus, dass die Verarbeitungsprozesse, die dem Verstehen von auditiv wahrgenommener sprachlicher Information zugrundeliegen, sich grundsätzlich von jenen unterscheiden, die dem Verständnis schriftlich dargebotener sprachlicher Information zugrundeliegen (Nickerson, 1981; Sachs, 1974; Spearritt, 1962; Danks & End, 1987; Samuels, 1987). Die Unterschiede werden in Leistungsunterschieden der entsprechenden modalitätsspezifischen kognitiven Fähigkeiten gesehen. Die Darbietungsmodalitäten stellen unterschiedliche Anforderungen an eine Person, z. B. beanspruchen sie das Gedächtnis auf unterschiedliche Weise. Darüber hinaus erfordert das Lesen stärker die Anwendung metakognitiver Strategien im Hinblick auf Organisation und Strukturierung, während das Hörver-

ständnis durch Sprechpausen und Intonation unterstützt wird (vgl. D. H. Rost & Buch, 2010; Danks & End, 1987; Samuels, 1987).

Für unterschiedliche Prozesse sprechen Befunde, die Unterschiede zwischen Hör- und Leseverständnis zeigen konnten. So fanden D. H. Rost und Hartmann (1992), dass Viertklässler bessere Hör- als Leseverständnisleistungen erbrachten, wobei sich jedoch – was dann wieder der monistischen Position entspricht – bei zunehmender Automatisierung des Leseprozesses die Leistungen anglichen. Außerdem zeigten sich kleine, jedoch kaum replizierbare, qualitative Unterschiede in der Wiedergabeleistung, die z. T. auf die Belastung des Verstehensprozesses oder Unterschiede zwischen den Versuchsgruppen zurückgeführt wurden (D. H. Rost & Buch, 2010).

4.4 Fazit

Zusammenfassend lässt sich sagen, dass sicheres Leseverständnis das Funktionieren und problemlose Zusammenspiel vielfältiger Teilprozesse erfordert. Für das Verstehen von Texten ist es zunächst nötig, dass die basalen Leseprozesse, also das Rekodieren und Dekodieren von Wörtern automatisiert sind, damit Kapazitäten für die Sinnerfassung frei werden. Die Sinnerfassung impliziert eine Text-Leser-Interaktion, wobei hierarchieniedrige Prozesse dem Wortverständnis und der Herstellung lokaler Kohärenzen dienen und hierarchiehohe Prozesse zum Aufbau globaler Kohärenzen beitragen.

Beim geübten Leser werden Textinformationen nicht nur oberflächlich repräsentiert und es werden nicht nur explizit im Text enthaltene Informationen zum Verständnis herangezogen, vielmehr werden die Informationen zueinander in Bezug gesetzt, verdichtet und zusammengefasst, zugleich aber auch elaboriert, durch Vor- und Weltwissen angereichert und dabei in die vorhandene Wissenstruktur integriert. Je nach Erfahrungen, Vorwissen und Erwartungen des Lesers läuft dieser Prozess unterschiedlich ab und es entsteht eine individuelle mentale Repräsentation, die nicht mehr genau dem Textinhalt entsprechen muss, und die als Situationsmodell bezeichnet wird. Während jüngere Kinder Texte zunächst vorwiegend oberflächlich repräsentieren, scheint im Alter von sieben bis elf Jahren ein Übergang zur vorwiegenden Repräsentation in Form von sogenannten „Fuzzy Traces“ stattzufinden (vgl. Nieding, 2006, S. 182). Schüler im Sekundarschulalter sollten demnach Textinhalte eher in Form von Fuzzy Traces speichern als in Form einer wörtlichen Repräsentation. Ob diese Fuzzy Traces eher textbasiert sind oder ein elaboriertes Situationsmodells darstellen, scheint schließlich vom Vorwissen abzuhängen (vgl. van den Broek & Gustafson, 1999). Über die oberflächen- und inhaltsbezogenen Ebenen hinaus werden als kognitiv anspruchsvollste Ebenen die Metaebenen des Erkennens von genrespezifischen Textstrukturen sowie von Darstellungsstrategien des Autors beschrieben. Diese sind insbesondere für eine angemessene Textinterpretation wichtig.

Trotz der vielfältigen am Lesen beteiligten Prozesse lassen sich empirisch in der Regel nur wenige Dimensionen voneinander abgrenzen. Während einige Autoren eine einzige Leseverständnisdimension postulieren (z. B. D. H. Rost & Buch, 2010; A. Voss, 2006), geht z. B. der Simple View of Reading von den zwei Dimensionen basale Lesetechnik und rezeptives Sprachverständnis aus (vgl. Hoover & Gough, 1990). Dabei scheint es auch von der Art der Aufgabenstellung abzuhängen, ob eine oder zwei Dimensionen voneinander abgegrenzt werden können. In neueren Studien zeigten sich Hinweise darauf, dass zusätzlich eine Geschwindigkeitskomponente eine Rolle spielen könnte (z. B. Carver, 1993; Joshi & Aaron, 2000; Kershaw & Schatschneider, 2012). Diesbezüglich ist jedoch noch weitere Forschung erforderlich.

Im Hinblick auf den Zusammenhang von Lese- und Hörverstehen ist sich die Forschung inzwischen weitgehend einig, dass die gefundenen hohen Korrelationen bei geübten Lesern darauf hindeuten, dass dem Lese- und dem Hörverstehen eine gemeinsame Verstehenskomponente zugrunde liegt. Unterschiede wurden vor allem bei Leseanfängern gefunden, bei welchen vermutlich eine unzureichende Dekodierfähigkeit die Verständnisleistung beim Lesen begrenzt. Bei geübten Lesern zeigen sich kaum noch Unterschiede.

Es lässt sich also schlussfolgern, dass dem aktuellen Forschungsstand zufolge in Bezug auf das Leseverständnis zwei Komponenten unterschieden werden können (vgl. Schneider, 2008): Die basale Lesekompetenz und das Leseverständnis. Beide Komponenten sind zwar nicht unabhängig voneinander, aber auch nicht sehr eng verknüpft. Das Leseverständnis scheint sich trotz seiner Komplexität und der zahlreichen Teilprozesse nicht in weitere distinkte Dimensionen unterteilen zu lassen. Zudem scheint es sich um modalitätsunabhängiges, generelles rezeptives Sprachverständnis zu handeln. Um nun zu erkennen, in Bezug auf welche Komponente bzw. welchen Teilprozess bei schwachen Leseleistungen im Einzelfall Defizite vorliegen und entsprechend angemessen fördern zu können, ist zunächst eine adäquate Diagnostik erforderlich. Mit dieser beschäftigt sich das nachfolgende Kapitel.

Kapitel 5

Diagnostik von Leseverständnis

Eine adäquate Beurteilung des aktuellen Leistungsstandes ist eine Voraussetzung für angemessene Unterrichtsgestaltung und gezielte Fördermaßnahmen. Bisweilen kann eine Diagnostik des Leseverständnisses auch für Entscheidungen bezüglich der weiteren Schul- und Berufslaufbahn hilfreich sein. Die Beurteilung des Leseverständnisses kann dabei auf unterschiedliche Weise geschehen. So kann die Lehrkraft die Leistung ihrer Schüler aufgrund von Unterrichtsbeobachtungen sowie schriftlichen oder mündlichen Prüfungen evaluieren und diese Evaluation in Form von Schulnoten, einer verbalen Leistungsbeurteilung oder eines Ratings ausdrücken. Weiter gibt es zahlreiche standardisierte Lesetests, die verschiedene Aspekte von Leseverständnis zuverlässig und objektiv erfassen. Für eine umfassende Beurteilung der Lesekompetenz können verschiedene Informationsquellen kombiniert werden.

Das vorliegende Kapitel betrachtet zunächst mögliche klinische Diagnosen, die bei Defiziten im Bereich des Lesens vergeben werden können, bevor es auf die Güte des Lehrerurteils zum Leseverständnis eingeht. Anschließend werden theoretische und methodische Grundlagen, Möglichkeiten und Grenzen sowie einige Beispiele standardisierter Lesetests dargestellt und diskutiert. Aufgrund der Relevanz für die vorliegende Arbeit liegt der Fokus dabei meist auf dem Sekundarschulalter.

5.1 Klinische Diagnosen

In Bezug auf Defizite im Bereich des Lesens existieren viele Begriffe (z. B. Dyslexie, Legasthenie, Lesestörung, Lese-Rechtschreibstörung), die teilweise nicht eindeutig bzw. nicht einheitlich definiert sind. Auf diese Begrifflichkeiten wird an dieser Stelle nicht im Detail eingegangen. Vielmehr werden die im Zusammenhang mit Schwierigkeiten im Bereich des Lesens möglichen klinischen Diagnosen kurz vorgestellt.

Gemäß dem aktuellen *Diagnostischen und Statistischen Manual psychischer Störungen* (DSM-5, American Psychiatric Association, 2013) kann bei dauerhaften Defiziten im Bereich des Lesens, die akademische oder berufliche Leistungen oder die alltägliche Lebensführung deutlich beeinträchtigen, eine „Spezifische Lernstörung“ mit Beeinträchtigungen im Lesen diagnostiziert werden. Dabei müssen die Leseleistungen deutlich unter dem liegen, was aufgrund des Alters, des gemessenen intellektuellen Niveaus und der altersgemäßen Bildung bei einer Person zu erwarten wäre. Entwick-

lungsstörungen, neurologische, sensorische oder motorische Störungen müssen als Ursache ausgeschlossen sein.

Da Lesen und Schreiben eng verknüpft sind, liegen häufig Defizite in beiden Bereichen vor. In diesen Fällen wird von einem gestörten Schriftspracherwerb gesprochen. In der *International Classification of Diseases* der Weltgesundheitsorganisation (ICD-10, Dilling, 2005) findet man diesbezüglich in der Kategorie „umschriebene Entwicklungsstörungen schulischer Fertigkeiten“ die „Lese-Rechtschreibstörung“. Diese besteht in einer im Vergleich zur Normstichprobe deutlich erhöhten Anzahl von Rechtschreibfehlern, Leseproblemen in Form vieler Lesefehler und stark verlangsamtem Lesetempo. Die Ursache für diese Probleme darf weder in einer geistigen Behinderung, einer Hör- oder Sehstörung oder einer neurologischen Krankheit noch in unzureichendem Unterricht liegen; zudem müssen die Schulleistungen deutlich beeinträchtigt sein.

Laut der sogenannten „Diskrepanzdefinition“ liegt dann eine Lese-Rechtschreibstörung vor, wenn zwischen allgemeinem intellektuellem Niveau und Lese-Rechtschreibleistung eine deutliche Diskrepanz besteht (H. Marx & Reinhold, 2010). Demnach werden Personen mit allgemein schwacher Lese- und Rechtschreibleistung ohne Diskrepanz zwischen Schriftsprachleistung und Intelligenzniveau als „allgemein lese-rechtschreibschwach“ bezeichnet, während Personen, bei denen eine entsprechende Diskrepanz vorliegt als „legasthen“ eingeordnet werden. Die Diskrepanzdefinition ist jedoch inzwischen umstritten, weshalb die neuere psychologische Forschung eher den deskriptiven Begriff „Lese-Rechtschreibschwierigkeiten“ (LRS) verwendet (K. E. Stanovich, Siegel & Gottardo, 1997; Klicpera & Gasteiger Klicpera, 2001; Pennington, Gilger, Olson & DeFries, 1992; P. Marx, Weber & Schneider, 2001; Metz, Marx, Weber & Schneider, 2003; H. Marx & Reinhold, 2010). Bei LRS spielt allerdings die Rechtschreibkompetenz eine größere Rolle als die Lesekompetenz, insbesondere in der Sekundarstufe ist die Lesekompetenz auch bei Schülern mit LRS recht gut entwickelt, während die Rechtschreibkompetenz weiter deutlich beeinträchtigt ist. Da das Lesen aber grundsätzlich tangiert ist und für die Validitätsanalysen in der vorliegenden Arbeit die LRS-Diagnose herangezogen wurde, sollte diese kurz erläutert werden.

Insgesamt sollten beim Verdacht auf eine Störung im Schriftsprachbereich in jedem Fall neben einer gründlichen Diagnostik der Lesekompetenz auch die Rechtschreibleistung und die Intelligenz sowie das Vorliegen komorbider Störungen geprüft werden (H. Marx & Reinhold, 2010). Darüber hinaus ist es notwendig, neurologische und sensorische Einschränkungen als Ursache für die schwachen Lese- und Rechtschreibleistungen auszuschließen (Schulte-Körne, 2004). Zudem sind Motorik, Artikulation und Sprachverständnis zu überprüfen und es sollte ein Bericht der Schule über den Schriftspracherwerb des betreffenden Schülers hinzugezogen werden. Die aus den verschiedenen Quellen zusammengetragenen Informationen sind für eine klinische Diagnose zu einem Gesamtbild zu integrieren.

5.2 Diagnostische Kompetenz von Lehrkräften

Als diagnostische Kompetenz wird die Fähigkeit von Lehrkräften bezeichnet, Schülerleistungen und Aufgabenschwierigkeiten korrekt einzuschätzen sowie zu wissen, welche Leistungen von Schülern einer bestimmten Klassenstufe und Schulart generell erwartbar sind (Artelt, 2009; Rjosk, McElvany, Anders & Becker, 2011; Schrader, 2010). Es handelt sich häufig um informelle Diagnostik in Form impliziter subjektiver Urteile, Einschätzungen und Erwartungen, die beiläufig und unsystematisch im Unterrichtsalltag gewonnen werden (Schrader, 2010).

Die Urteilsgenauigkeit wird durch Merkmale sowohl der Lehrkraft als auch der Klassen und Schüler beeinflusst (Rjosk et al., 2011). Große Unterschiede zwischen Studien einerseits sowie zwischen Lehrkräften und Klassen in derselben Studie andererseits werden zum einen auf Einflüsse der Kontextbedingungen (z. B. Klassengröße, klasseninterne Leistungsheterogenität) zurückgeführt und zum anderen auf tatsächliche Unterschiede in der diagnostischen Kompetenz der Lehrkräfte (vgl. Schrader, 2010; Artelt, 2009). Lehrkräfte haben bei der Leistungsbeurteilung oft nur die einzelne Klasse oder Schule als Referenzmaß (vgl. z. B. Artelt, 2009). Typische Beobachtungsverzerrungen (z. B. selektive Wahrnehmung, Erinnerungsfehler, Erwartungseffekte), die aktuelle Stimmungslage sowie Sympathien und Antipathien kommen hinzu (Tent, 2006). Nicht zuletzt ist beim Lehrerurteil zu berücksichtigen, dass die Lehrkraft dabei indirekt auch ihre eigene Leistung bewertet. Aus diesen Gründen wird das Lehrerurteil häufig als subjektiv, unreliabel und wenig valide kritisiert (z. B. Schrader, 2010). Allerdings können Lehrerurteile valider sein als standardisierte Tests. Die logische Validität von Schulnoten z. B. resultiert schon daraus, dass Schulerfolg in erster Linie über sie definiert wird (Tent, 2006). Dennoch ist es problematisch, dass die Güte des Lehrerurteils nicht offensichtlich ist und daher bei der Interpretation nicht berücksichtigt werden kann. Die große Bedeutung, die dem Lehrerurteil z. B. bei weitreichenden Schullaufbahnentscheidungen zukommt, lässt eine empirische Validierung daher notwendig erscheinen.

Die Güte von Lehrerurteilen wird meist anhand klassenspezifischer Korrelationen von Lehrerurteilen und Schülerleistungen bestimmt (Karing, 2009; Tent, 2006). Das Lehrerurteil besteht in der Regel aus der fachspezifischen Schulnote oder einem Rating der Schülerleistung durch die Lehrkräfte, die Schülerleistung wird über standardisierte Leistungstests erfasst. Auf diese Weise wird die Rangkomponente bzw. der soziale Vergleich zur Bestimmung der Urteilsgüte herangezogen (vgl. Karing, 2009; Spinath, 2005; Rjosk et al., 2011). Das absolute Leistungsniveau wird nicht berücksichtigt, Über- oder Unterschätzungen der gesamten Klasse bleiben dadurch unbemerkt (vgl. Karing et al., 2011; Spinath, 2005; Südkamp & Möller, 2009). Die Korrelationen zwischen Schülerleistung und Lehrerurteil im Hinblick auf verschiedene Leistungsbereiche belaufen sich laut einer Metaanalyse von Südkamp, Kaiser und Möller (2012) auf etwa $r = .63$. Lehrkräfte scheinen Schülerleistungen somit im Allgemeinen recht

akkurat einschätzen zu können (vgl. z. B. Helmke, 2004; Hoge & Coladarci, 1989; Rjosk et al., 2011). Bei PISA streuten jedoch die Schulnoten von Schülern, die sich objektiven Leistungstests zufolge auf der gleichen Niveaustufe befanden, über mehrere Notenstufen. Sogar innerhalb eines Bundeslandes und derselben Schulform wurden die Noten zum Teil nach sehr unterschiedlichen Maßstäben vergeben (Artelt, 2009; Baumert, Trautwein & Artelt, 2003; Bos et al., 2003). Eine perfekte Übereinstimmung von Lehrerurteil und Testergebnis ist allerdings schon deshalb nicht zu erwarten, weil ein Testergebnis eine Art Momentaufnahme darstellt, während sich das Lehrerurteil auf alltägliche Unterrichtsbeobachtungen sowie mehrere schriftliche und mündliche Leistungserhebungen beziehen kann. Dass die Korrelationen zwischen Schulnoten und Testergebnissen in der Regel mittelhoch ausfallen, könnte zudem so interpretiert werden, dass Schulnoten zum einen die Leistungen nicht vollständig abbilden und zum anderen über die objektive Leistung hinaus weitere Informationen enthalten, die z. B. auf systematische Urteilsfehler zurückzuführen sind (Tent, 2006).

Beim Vergleich des Lehrerurteils mit objektiven Testergebnissen ist es zudem wichtig, die Testgüte und die dem Test zugrunde liegende Konzeption vom interessierenden Merkmal zu berücksichtigen (z. B. welche Aspekte von Lesekompetenz ein Test erfasst und welche nicht). Nicht zuletzt sollte bei der Beurteilung der diagnostischen Kompetenz von Lehrkräften – also bei der Korrelation der Lehrerurteile mit den Ergebnissen standardisierter Tests – bedacht werden, dass umgekehrt zur Validierung von standardisierten Tests häufig Schulnoten oder Lehrerratings verwendet werden. Somit beißt sich die Katze in den Schwanz, wenn einerseits zur Beurteilung der diagnostischen Kompetenz standardisierte Tests und andererseits zur Beurteilung der Testgüte das Lehrerurteil herangezogen werden. Dieser theoretisch begründete Validierungszirkel ist jedoch laut Tent (2006) derzeit alternativlos und entspricht dem Prinzip der Methodenkonvergenz.

Beurteilung der Leseleistungen. Studien aus dem englischen Sprachraum berichten Korrelationen von $r = .62$ (Bates & Nettelbeck, 2001) bis $r = .82$ (Demaray & Elliot, 1998) zwischen Lehrerurteil zur Lesefertigkeit bzw. zum flüssigen Vorlesen und den Ergebnissen entsprechender standardisierter Tests. Auch in Bezug auf das Lesen sind die Werte somit akzeptabel (vgl. Bates & Nettelbeck, 2001; Demaray & Elliot, 1998; Lorenz & Artelt, 2009; Südkamp & Möller, 2009; Karing et al., 2011). Allerdings streut die Urteilsgenauigkeit zwischen den Lehrkräften stark, und nur wenige Lehrkräfte können aus Vorlesefehlern Rückschlüsse auf verschiedene Arten von Leseschwierigkeiten ziehen (z. B. Artelt, 2009; Hopkins, George & Williams, 1985; Spinner, 2004). Der Großteil der Befunde zur diagnostischen Kompetenz von Lehrkräften bezüglich der Lesekompetenz basiert auf Grundschuldaten (Bates & Nettelbeck, 2001; Coladarci, 1986; Demaray & Elliot, 1998; Feinberg & Shapiro, 2003; Hoge & Coladarci, 1989). Im Folgenden werden jedoch aufgrund der Relevanz für die vorliegende Arbeit nur Befunde zur Sekundarstufe dargestellt.

Karing (2009) verglich die diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften und fand, dass letztere u. a. das Textverständnis ihrer Schüler schlechter einschätzen konnten. Dies passt zu Befunden aus anderen Studien; Grundschullehrkräfte zeigten stets eine bessere diagnostische Kompetenz bezüglich der Lesekompetenz als Lehrkräfte weiterführender Schulen (z. B. Karing et al., 2011; Lorenz & Artelt, 2009; Bates & Nettelbeck, 2001). Dafür werden verschiedene (sich nicht gegenseitig ausschließende) Erklärungsmöglichkeiten angeführt: (1) Gymnasialklassen sind leistungshomogener, und kleinere Leistungsunterschiede lassen sich schwerer erkennen. (2) Am Gymnasium herrscht das Fachlehrerprinzip, im Gegensatz zum Klassenlehrerprinzip der Grundschule, was dazu führt, dass Grundschullehrkräfte die Schüler besser kennen. (3) Die Lehrerausbildung ist für das Gymnasium stärker auf das Fachwissen fokussiert, während im Grundschulbereich pädagogisch-didaktische Lehrveranstaltungen stärker vertreten sind. (4) Explizite Leseförderung ist vor allem in der Grundschule wesentlicher Bestandteil des Deutschcurriculums und findet in der Sekundarstufe nur noch eingeschränkt statt (vgl. Streblow & Möller, 2010), weshalb die Lesekompetenz in den Schulleistungen der mittleren und höheren Klassenstufen nur indirekt zum Ausdruck kommt.

Bei Karing et al. (2011) korrelierten Lehrerurteile im Bereich Lesen in der Sekundarstufe I sowohl auf Konstrukt- als auch auf Aufgabenebene nur schwach mit den Schülerleistungen. Globale Einschätzungen fielen dabei im Vergleich zu aufgabenspezifischen Beurteilungen akkurater aus. Auf Aufgabenebene überschätzten die Deutschlehrkräfte das Lesekompetenzniveau im Mittel konform mit Ergebnissen anderer Studien (z. B. Schrader, 1989; Schrader & Helmke, 1987; Südkamp & Möller, 2009). Auch Rjosk et al. (2011) kommen bei ihrer Untersuchung der diagnostischen Kompetenz von Deutschlehrkräften hinsichtlich der basalen Lesekompetenz von Sechstklässlern zu dem Ergebnis, dass die Lehrkräfte das Niveau ihrer Schüler überschätzten. Wie zuvor bereits in anderen Studien (z. B. Bos et al., 2003) zeigte sich zudem, dass die Lehrkräfte interindividuell deutlich in ihrer Einschätzung differieren, was als Hinweis auf das Fehlen eines allgemein gültigen Maßstabs gewertet wird. Besonders hervorgehoben sei an dieser Stelle auch ein Befund der nationalen Zusatzuntersuchung von PISA 2000: die Lehrkräfte der Hauptschulen waren in dieser Untersuchung nur sehr eingeschränkt in der Lage, Risikoschüler als solche zu identifizieren (Artelt et al., 2001, S. 119). Nur 11.4 % der Schüler unter Niveaustufe I wurden als schwache Leser erkannt. Natürlich misst auch der PISA-Lesetest nicht perfekt, jedoch besteht er aus Aufgaben mit hoher Alltagsrelevanz, weshalb die geringe Übereinstimmung von Test und Lehrerurteil durchaus bedenklich ist (vgl. W. Lenhard, 2013, S. 66).

5.3 Standardisierte Lesetests

Aufgrund der im vergangenen Abschnitt dargestellten Probleme der Sekundarschullehrkräfte, das Leseverständnis ihrer Schüler adäquat zu beurteilen, sollten für Entscheidungen von größerer Tragweite zur Absicherung ergänzend zuverlässige, valide und standardisierte Tests zum Einsatz kommen (Tent, 2006; Schrader, 2010). Standardisierte Tests ermöglichen eine fairere Leistungsbeurteilung sowie eine objektivere Überprüfung der Lernzielerreichung. Weiter können sie Lehrkräfte bei der Unterrichtsplanung und bei der Auswahl von Unterrichtsmaterialien unterstützen. Auch für die Individualdiagnostik (z. B. LRS-Diagnostik), zu Forschungszwecken (z. B. Evaluation von Fördermaßnahmen, Lerneinheiten und Unterrichtsmethoden), im Rahmen größerer nationaler und internationaler Vergleichsstudien sowie zur Einschätzung der diagnostischen Kompetenz von Lehrkräften können standardisierte Tests eingesetzt werden. Sie tragen zu einer adäquaten Diagnose, Förderung und Prognose bei (Rjosk et al., 2011).

Im Folgenden werden zunächst testtheoretische Grundlagen dargestellt, die zum Verständnis der psychometrischen Diagnostik von Leseverständnis nötig sind. Anschließend werden Gütekriterien zur Bewertung von Tests besprochen. Darauf folgen eine Darstellung verschiedener Möglichkeiten zur Operationalisierung von (Teil-)Prozessen des Lesens, eine Diskussion grundsätzlicher Möglichkeiten und Grenzen standardisierter Tests sowie eine Beschreibung und kritische Betrachtung verfügbarer Lesetests. Schließlich werden die PISA-Lesetests genauer beleuchtet, da die Befunde der PISA-Studien und somit die mit diesen Tests erhobenen Daten für die vorliegende Arbeit von besonderer Bedeutung sind.

5.3.1 Testtheorie

Psychometrische Lesetests sind wissenschaftliche Routineverfahren zur Untersuchung eines empirisch abgrenzbaren Merkmals (z. B. Lesegeschwindigkeit oder Textverständnis) mit dem Ziel einer möglichst quantitativen Aussage über einen relativen Grad der individuellen Merkmalsausprägung (vgl. Lienert & Raatz, 1998, S. 1). Psychometrische Tests zeichnen sich dadurch aus, dass sie gemäß der „Klassischen Testtheorie“ und/oder der „Item Response Theorie“ (Erklärungen s. Kap. 5.3.1.1 und 5.3.1.2) konstruiert werden und genau definierten Gütekriterien entsprechen (vgl. Bühner, 2011, S. 38). Die Gütekriterien stellen sicher, dass die Tests zuverlässige und faire Beurteilungen ermöglichen. Die zugrunde gelegte Theorie erlaubt es, die Kriterien zu überprüfen und gegebenenfalls ungeeignete Items zu eliminieren. Ungeeignete Items können in der Praxis schwerwiegende Fehlentscheidungen z. B. für die Schullaufbahn oder den Förderbedarf nach sich ziehen. Erst wenn ein Test die Gütekriterien erfüllt, ist eine sinnvolle Interpretation der Testergebnisse möglich und der Test kann in der Praxis eingesetzt werden. Daher ist eine sorgfältige Testkonstruktion unerlässlich.

Im Rahmen der psychometrischen Testkonstruktion bezieht sich der Begriff „Testtheorie“ auf die dem Test zugrunde liegende Theorie über den Zusammenhang zwischen dem von einer Person in der Testsituation gezeigten Verhalten und dem zu messenden psychischen Merkmal (J. Rost, 2004, S. 21). Dabei geht es darum, wie sich das interessierende Merkmal der Personen auf ihr Verhalten in der Testsituation, also z. B. die Reaktionen auf die Testitems, auswirkt. Eine solche Theorie ist deshalb wichtig, um umgekehrt aus dem Testergebnis auf das psychische Merkmal der Person schließen zu können.

Da Lesetests die Leseleistung messen, gehören sie zur Kategorie der sogenannten „Leistungstests“. Das Messen einer Leistung setzt voraus, dass diese nach bestimmten Kriterien als erbracht oder nicht erbracht beurteilt werden kann (Bühner, 2011, S. 21). Da das Leseverständnis einer Person nicht direkt beobachtbar ist, muss es aus beobachtbarem Verhalten erschlossen werden. Es handelt sich also um ein Konstrukt. Die Testitems sind die manifesten (beobachtbaren) Variablen, die als Indikatoren für das Konstrukt bzw. die latente (nicht direkt beobachtbare) Variable herangezogen werden. Während ein Konstrukt keine mess- oder testtheoretischen Eigenschaften voraussetzt, bezieht sich der Begriff „latente Variable“ meist auf genau eine Variable (J. Rost, 2004, S. 30).

Ein Konstrukt wird also über mehrere verschiedene Items gemessen, wobei die Reaktionen der Testperson auf die Items das beobachtbare Verhalten darstellen, welches z. B. durch das Ankreuzen einer Antwortalternative zunächst nur zählbar gemacht wird (Bühner, 2011, S. 31f.). Nur sofern die Zuordnung der Zahlen zu Personenmerkmalen die Relationen des Merkmals zwischen den Personen abbildet, kann auch von „messen“ gesprochen werden. Damit ein auffälliges Testergebnis eindeutig auf eine extreme Ausprägung der zu messenden, latenten Variable zurückgeführt werden kann und nicht mehrere Erklärungen denkbar sind, muss das zu messende Konstrukt eindimensional sein. Unterschiede in den Reaktionen der Testpersonen auf die Items sollten ausschließlich auf Unterschiede der Personen in der latenten Variable zurückzuführen und somit für die Korrelation der Itemwerte verantwortlich sein. Weitere Variablen dürfen nicht für Unterschiede der Reaktionen der Testpersonen auf die Items verantwortlich sein. Wird die latente Variable konstant gehalten, müssen die Itemreaktionen somit unkorreliert sein, was auch als „lokale Unabhängigkeit“ bezeichnet wird.

Leistungstests werden im Allgemeinen noch weiter in Speedtests und Niveautests unterteilt (vgl. z. B. Amelang & Schmidt-Atzert, 2006; J. Rost, 2004; Bühner, 2011). Beiden gemeinsam ist, dass das Ergebnis von der Versuchsperson nur nach unten verfälschbar ist, aber nicht nach oben (J. Rost, 2004, S. 43). Der Unterschied liegt darin, dass bei Speedtests der Schwerpunkt auf der Bearbeitungsgeschwindigkeit liegt, während Leistungstests die maximal mögliche Leistung bei unbegrenzter Bearbeitungszeit erfassen.

Speedtests bestehen in der Regel aus leichten bis mittelschweren Items (z. B. Wörter oder einfache Sätze), die ohne im Voraus festgelegte Begrenzung der Bearbeitungs-

dauer der einzelnen Items, einzelner Itemgruppen oder des gesamten Tests von allen Personen der Zielgruppe gelöst (z. B. korrekt gelesen) werden können. Ziel ist es, die Leistung in Form der Bearbeitungsgeschwindigkeit (z. B. Lesegeschwindigkeit) zu messen. Dafür gibt es zwei Möglichkeiten: (1) Die Vorgabe einer Zeitbegrenzung, innerhalb derer (fast) kein Proband alle Items lösen kann (als Testergebnis wird die Anzahl der bearbeiteten Items oder die Anzahl der korrekt gelösten Items betrachtet) oder (2) die Vorgabe einer bestimmten Anzahl an Items und Stoppen der Zeit, die eine Person für die Lösung jedes einzelnen Items oder des gesamten Tests benötigt (das Testergebnis besteht in der Bearbeitungszeit).

Die Bearbeitung soll bei allen Speedtests explizit so schnell und gleichzeitig so genau wie möglich erfolgen. Die Probanden machen dabei natürlich Fehler, die dann entweder als separater Kennwert berücksichtigt werden oder in die Berechnung des Testergebnisses einfließen (das dann z. B. in der Anzahl der korrekt gelesenen Wörter besteht). Für viele Speedtests ist kein Testmodell nötig, denn die für die Lösung einer bestimmten Itemanzahl benötigte Zeitdauer stellt bereits eine metrische Personenvariable dar (J. Rost, 2004, S. 44). Andererseits kann bei Festlegung einer bestimmten Zeitdauer, die Anzahl richtig gelöster Items als metrische Personenvariable aufgefasst werden.

Bei *Niveautests* steht demgegenüber das korrekte Lösen von Items im Vordergrund. Bei der Erfassung der maximalen Leistung ist die Bearbeitungsgeschwindigkeit von untergeordneter Bedeutung. Die Items (z. B. Verständnisfragen zu einem Text) sind in der Regel unterschiedlich schwer und meist nach Schwierigkeitsgrad aufsteigend angeordnet. Die schwersten Items können selbst bei unbegrenzter Bearbeitungszeit nur von den wenigsten Testpersonen korrekt gelöst werden. Die Bearbeitungszeit wird nicht oder nur großzügig begrenzt. Reine Niveautests sind schon aus technischen Gründen nicht umsetzbar, da kein Test unendlich lang vorgegeben werden kann. Die Zeitbegrenzung sollte daher so vorgegeben werden, dass alle Testpersonen alle Items bearbeiten können (J. Rost, 2004, S. 43).

In den folgenden Abschnitten werden die beiden den meisten psychometrischen Tests zugrunde liegenden Testtheorien – die Klassische Testtheorie und die Item Response Theorie – vorgestellt.

5.3.1.1 Klassische Testtheorie

Die Klassische Testtheorie (KTT) ermöglicht es, unter bestimmten Vorannahmen die Messgenauigkeit (Reliabilität) eines Tests zu bestimmen (Bühner, 2011, S. 41). Bei mehreren Messungen ergeben sich in der Regel für ein und dieselbe Person unterschiedliche Testergebnisse. Dies kann auf Übungseffekte, Transfereffekte oder unsystematische (äußere oder innere) Einflüsse – meist auf eine Kombination aus allen drei – zurückzuführen sein. Die unsystematischen Einflüsse führen zu verzerrten Mes-

sungen und bilden den Messfehler. Die KTT berücksichtigt nur die unsystematischen Fehler und nimmt an, dass diese zufällig und normalverteilt sind.

Die KTT nach Lord und Novick (1968, S. 55ff.) basiert auf drei Axiomen, aus denen deduktiv weitere Aussagen abgeleitet werden können. Das erste Axiom besagt, dass der „wahre Wert“ des zu messenden Merkmals einer Person gleich dem Erwartungswert des Messwertes dieser Person bei diesem Test ist. Das zweite Axiom besagt, dass sich der Messwert (X) einer Person bei einem Test zusammensetzt aus dem wahren Wert (T) dieser Person bei diesem Test und dem Messfehler (E). Entsprechend lautet die Grundgleichung:

$$X = T + E$$

Aus den ersten beiden Axiomen resultiert, dass der Erwartungswert des Messfehlers Null ist und dass die Varianz des Messwerts einer Person bei einem Test lediglich aus der Varianz des Messfehlers besteht. Das dritte Axiom besagt, dass die Korrelation zwischen den Messfehlern einer Person bei zwei Tests gleich Null ist.

Der Messwert wird in der KTT meist aus der Summe der Itemantworten gebildet und als Summenscore bezeichnet. Bei Gültigkeit der Axiome ist der Summenscore ein erwartungstreuer, unverzerrter Schätzer des wahren Wertes und damit des zu messenden Merkmals.

Unter bestimmten Annahmen kann mithilfe der KTT zum einen die Reliabilität eines Tests bestimmt werden, zum anderen kann über die Bestimmung und Berücksichtigung des Messfehlers ausgehend vom Testergebnis einer Person ein Bereich bestimmt werden, innerhalb dessen der wahre Wert dieser Person mit einer bestimmten Wahrscheinlichkeit liegt.

Testkonstruktion gemäß der KTT. Der erste Schritt der Testkonstruktion besteht darin, das interessierende Merkmal theoriegeleitet so genau wie möglich zu definieren (Fisseni, 2004, S. 28). Anschließend sind Items aus der Theorie abzuleiten. Die Items sind dann zu einer empirischen Erprobung einer Stichprobe vorzulegen, die der Zielgruppe möglichst ähnlich ist. Auf der Grundlage dieser Daten kann eine Itemanalyse durchgeführt werden, indem statistische Kennwerte zur Güte der Items ermittelt werden. Meist werden die Trennschärfe, der Schwierigkeitsindex und die Homogenität ermittelt.

Die *Trennschärfe* (r_{it}) besteht in der Korrelation des Itemscores mit dem Summenscore (Bortz & Döring, 2006, S. 219). In der Regel wird eine (part-whole-) korrigierte Trennschärfe berechnet, d. h. der Wert des interessierenden Items wird mit dem Testergebnis korreliert, wobei der Wert des interessierenden Items zuvor vom Testergebnis subtrahiert wird, um eine partielle Eigenkorrelation und eine damit einhergehende Überschätzung der Trennschärfe zu vermeiden (Moosbrugger, 2008a). Da es sich bei der Trennschärfe um eine Korrelation handelt, nimmt der Koeffizient Werte zwischen -1 und 1 an. Ein hoher positiver Wert für die Trennschärfe drückt aus, dass insgesamt leistungsstarke Schüler das Item meistens lösen und insgesamt leistungsschwa-

che Schüler das Item meistens nicht lösen. Eine negative Trennschärfe bedeutet demgegenüber, dass ein Item von insgesamt eher leistungsschwachen Schülern häufiger gelöst wird als von insgesamt eher leistungsstarken Schülern. Items mit negativen Trennschärfen sind daher auf jeden Fall zu eliminieren. Bei der KTT werden hohe Trennschärfen angestrebt, um eine hohe Messgenauigkeit des Tests zu erreichen. Die *Itemschwierigkeit* (p) wird als der prozentuale Anteil korrekter Lösungen eines Items in der Stichprobe verstanden (Lienert & Raatz, 1998, S. 73). Ein höherer Schwierigkeitsindex steht somit kontraintuitiv für einen höheren Lösungsanteil. Bei der KTT werden Items mit Schwierigkeitswerten zwischen .20 und .80 angestrebt, denn mittlere Schwierigkeitswerte sind eine Voraussetzung für hohe Trennschärfen. Mit *Homogenität* ist der Bereich der inhaltlichen Überlappung der Items gemeint, der möglichst groß sein sollte, da alle Items dasselbe Merkmal (wenn auch jeweils eine andere Facette des Merkmals) erfassen sollen (Fisseni, 2004, S. 40). Im Rahmen der KTT wird die Größe des Überlappungsbereiches meist mithilfe von Item-Interkorrelationen oder Faktorenanalysen bestimmt.

Im nächsten Schritt folgt die Itemselektion. Items, die sich bei der Itemanalyse als ungeeignet (z. B. zu schwer, zu leicht oder zu wenig trennscharf) erwiesen haben, werden eliminiert. Über die genannten Itemkennwerte hinaus sind im Einzelfall weitere Aspekte (z. B. Art des Tests, Stichprobenzusammensetzung, Breite des Merkmalsausschnitts, inhaltliche Bedeutung eines Items für das Konstrukt) zu berücksichtigen (Bühner, 2011, S. 81). Somit ist die Itemselektion ein relativ subjektives Verfahren, da beispielsweise die Gewichtung der verschiedenen Aspekte (z. B. statistische Kennwerte vs. inhaltliche Bedeutsamkeit für das Konstrukt) in der Hand des Testkonstrukteurs liegt.

5.3.1.2 Item Response Theorie

Die Item Response Theorie (IRT) beschäftigt sich mit dem Zusammenhang zwischen Reaktionen von Testpersonen auf Items und Merkmalen der Testpersonen und der Items. Sie umfasst zahlreiche Testmodelle, die bestimmte Annahmen darüber machen, von welchen Modellparametern die Lösungswahrscheinlichkeit einer Testperson für ein Item abhängt (Bühner, 2011, S. 494). Wenn die Modellparameter bekannt sind, ist anhand der IRT – im Gegensatz zur KTT – über die Bestimmung der Lösungswahrscheinlichkeit einer Testperson für ein Item eine konkrete Verhaltensvorhersage möglich. Da die IRT Wahrscheinlichkeitsaussagen macht, wird sie auch als probabilistische Testtheorie bezeichnet.

Der erste Modellparameter besteht bei allen IRT-Modellen in der Fähigkeitsausprägung der Person (Personenparameter). Hinzu kommen je nach Modell weitere Parameter, z. B. beim „Ein-Parameter-Logistischen- (1-PL-) Modell“ die Itemschwierigkeit (Itemparameter), beim „Zwei-Parameter-Logistischen- (2-PL-) Modell“ über das 1-PL-Modell hinaus die Ratewahrscheinlichkeit, beim „Drei-Parameter-Logistischen- (3-PL-)

Modell“ über das 2-PL-Modell hinaus die Itemtrennschärfe oder beim „Mixed-Rasch-Modell“ die Zugehörigkeit einer Person zu einer bestimmten Personengruppe. Besonders wichtig ist das 1-PL-Modell, das meist – und auch im Folgenden – als „Rasch-Modell“ bezeichnet wird, obwohl es kleinere Unterschiede zwischen diesen beiden Modellen gibt (Bühner, 2011, S. 495).

Das Rasch-Modell. Das Rasch-Modell besagt, dass mit zunehmender Personenfähigkeit die Wahrscheinlichkeit für die Lösung eines Items steigt (Bühner, 2011, S. 57). Die Lösungswahrscheinlichkeit ist dabei von zwei Parametern abhängig: der Fähigkeit der Person (Personenparameter) und der Itemschwierigkeit (Itemparameter). Der Personenparameter gibt den Wert der latenten Variable für eine Person an, daher erhalten Personenparameter und latente Variable die gleiche Bezeichnung (Θ, Theta). Personenparameter und Itemschwierigkeit werden in der gleichen Einheit – der sogenannten „Logiteinheit“ – gemessen. Der entsprechende Wertebereich erstreckt sich theoretisch sowohl in negative als auch in positive Richtung unendlich, liegt üblicherweise jedoch im Bereich zwischen -3 und 3 (Bühner, 2011, S. 496). Negative Werte stehen für leichte Items und geringe Personenfähigkeiten, positive Werte für schwere Items und eine hohe Personenfähigkeit. Beide Parameter weisen mindestens Intervallniveau auf.

Da die Beziehung zwischen den Parametern probabilistisch ist, wird auch eine Person mit geringer Fähigkeit im Verhältnis zur Itemschwierigkeit das Item mit einer gewissen (geringen) Wahrscheinlichkeit lösen (Bühner, 2011, S. 492). Es wird genauer ein logistischer Zusammenhang zwischen der Differenz aus Item- und Personenparameter sowie der Lösungswahrscheinlichkeit für ein Item angenommen. D. h. die Wahrscheinlichkeit der Lösung des Items nimmt zu, je weiter die Personenfähigkeit die Itemschwierigkeit übersteigt. Ausgangsgleichung der Skalierung ist somit die Logistische Funktion, die den probabilistischen Zusammenhang zwischen Personenfähigkeit (Θ) und Itemschwierigkeit (σ) abbildet (Fisseni, 2004, S. 102):

$$p(x) = \frac{e^{(\Theta-\sigma)}}{1 + e^{(\Theta-\sigma)}}$$

Abbildung 3 veranschaulicht den probabilistischen Zusammenhang zwischen den Parametern anhand einer sogenannten „Item Characteristic Curve“ (ICC) für ein Item. Auf der x-Achse ist der Wert der latenten Variable aufgetragen, auf der Y-Achse die Lösungswahrscheinlichkeit. Mit zunehmender Personenfähigkeit steigt die Wahrscheinlichkeit, das Item korrekt zu lösen, kontinuierlich an. Die Steigung der Kurve an ihrer steilsten Stelle entspricht der Itemtrennschärfe, da an dieser Stelle Personen mit relativ ähnlichen Merkmalsausprägungen maximal unterschiedliche Wahrscheinlichkeiten haben, das Item zu lösen.

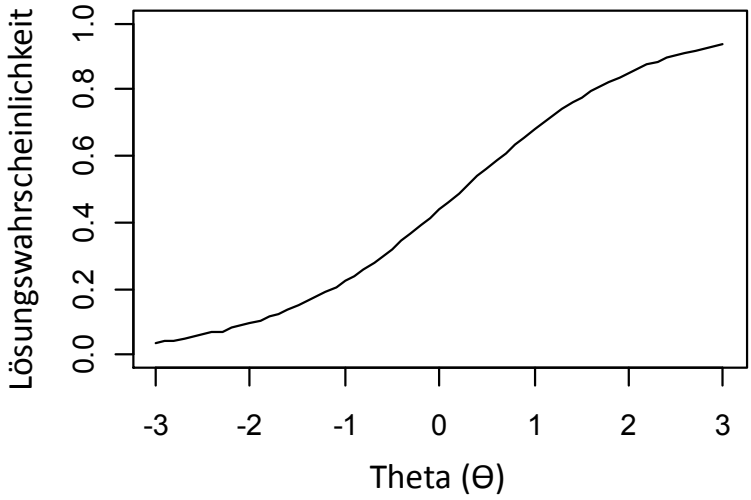


Abbildung 3. Beispiel einer ICC im Rasch-Modell.

Eine Voraussetzung dafür, dass es möglich ist, die direkte Differenz von Personen- und Itemparameter zu bilden, ist die *Eindimensionalität* der Skala. Diese besagt, dass Personen- und Itemparameter auf einer gemeinsamen latenten Dimension liegen (Strobl, 2010, S. 23). Eindimensionalität bedeutet, dass nur ein Merkmal gemessen wird und nicht gleichzeitig noch ein weiteres.

Im Rasch-Modell hängt die Lösungswahrscheinlichkeit für ein Item ausschließlich von der Personenfähigkeit und der Itemschwierigkeit ab, daher muss *Gleichheit der Trennschärfen* für alle Items gegeben sein. Gleichheit der Trennschärfen bedeutet, dass sich die ICCs der Items eines Rasch-Modell-konformen Tests nicht überschneiden. Leichtere Items liegen weiter links auf der x-Achse, schwerere Items weiter rechts, die Steigungen im Wendepunkt (also die Trennschärfen) sind jedoch für alle Items gleich. Zudem sollte sich die ICC gegen minus unendlich der x-Achse asymptotisch annähern. Geschieht dies nicht, können auch Personen mit einer sehr niedrigen Fähigkeit das Item mit einer unverhältnismäßig großen Wahrscheinlichkeit lösen. Dies ist ein Hinweis darauf, dass möglicherweise Raten eine Rolle spielt und dass die Berücksichtigung eines entsprechenden Parameters in Erwägung gezogen und somit vom Rasch-Modell abgewichen werden sollte oder entsprechende Items zu eliminieren sind.

Weiter gilt das Rasch-Modell nur, wenn *lokale stochastische Unabhängigkeit* vorliegt, d. h. die Lösungswahrscheinlichkeit für ein Item sollte unabhängig von der Lösungswahrscheinlichkeit eines anderen Items sein (Strobl, 2010, S. 18). Genauso sollte die Lösungswahrscheinlichkeit einer Person für ein Item unabhängig von der Lösungswahrscheinlichkeit einer anderen Person für dasselbe Item sein. „Lokal“ bedeutet hierbei, dass sich die Unabhängigkeit der Aufgaben nur auf eine Person oder mehrere Personen mit gleicher Fähigkeit bezieht. Eine Person mit höherer Fähigkeit kann selbst-

verständlich alle Aufgaben mit einer höheren Wahrscheinlichkeit lösen als eine Person mit geringerer Fähigkeit. Lokale stochastische Unabhängigkeit impliziert, dass die Items homogen sind (Bühner, 2011, S. 485). Dies ist eine notwendige, jedoch keine hinreichende Bedingung für die Eindimensionalität des Tests.

Sind die Itemtrennschärfen gleich und ist das Kriterium der lokalen stochastischen Unabhängigkeit der Items erfüllt, ist die Grundlage für *spezifisch objektive Vergleiche* von Items und Personen (Stichprobeninvarianz) gegeben. Spezifische Objektivität bedeutet, dass Differenzen zwischen Itemparametern unabhängig von der herangezogenen Personenstichprobe immer gleich sind ebenso wie Differenzen zwischen Personenparametern unabhängig von der verwendeten Itemstichprobe bis auf die Messfehler immer gleich sind (Bühner, 2011, S. 500). Gilt das Rasch-Modell, messen die Items also in jeder beliebigen Substichprobe von Personen dasselbe Merkmal (Fähigkeit oder Eigenschaft). Deshalb sollten die Schätzungen der Itemparameter für jede beliebige Substichprobe von Personen gleich ausfallen. Da somit bei allen Personen ein homogenes Merkmal gemessen wird, spricht man auch von „Personenhomogenität“ (Bühner, 2011, S. 539).

Ist eine Skala eindimensional und ihre Items sind konform mit dem Rasch-Modell (d. h. sie weisen gleiche Trennschärfen auf, die ICCs nähern sich in Richtung minus unendlich asymptotisch der x-Achse an und die Items sind lokal stochastisch unabhängig voneinander), dann ist der Summenscore der Items eine *suffiziente Statistik*. In diesem Fall sagt der Summenwert der Itemantworten tatsächlich etwas über den Ausprägungsgrad einer Person auf der latenten Variable aus (Bühner, 2011, S. 57). Dadurch wird die genauere Betrachtung des Antwortmusters überflüssig, da sie keine zusätzlichen Informationen über die Merkmalsausprägung liefert. Darüber hinaus sind die Messungen bei Gültigkeit des Rasch-Modells mindestens intervallskaliert (Strobl, 2010, S. 25). D. h. es ist ein metrisches Skalenniveau mit gleichabständigen Skalenabschnitten gegeben, was die Durchführung weiterer Berechnungen (z. B. Differenzen, Mittelwertbildung) erlaubt.

5.3.1.3 Kritische Bewertung der Theorien

Die beiden Testtheorien wurden lange Zeit als konkurrierende Ansätze der Testkonstruktion betrachtet. Dabei schließen sie sich aber nicht gegenseitig aus und gelten heute als sich komplementär ergänzend (vgl. z. B. J. Rost, 2004, S. 12). Die KTT fokussiert als Messfehlertheorie stärker auf die Reliabilität und ist für Anwendungsfragestellungen bezüglich der Kriteriumsvalidität (Erläuterung s. Kap. 5.3.2.1) ausreichend (Moosbrugger, 2008a). Der Schwerpunkt der IRT liegt dagegen auf der Skalierbarkeit und der Konstruktvalidität (Erläuterung s. Kap. 5.3.2.1). Die IRT ermöglicht im Gegensatz zur KTT eine für inhaltlich-theoretische Fragestellungen der Konstruktvalidität ausreichende Überprüfung der Eindimensionalität sowie der Item- und Personenhomogenität.

KTT. Die KTT ist eine reine Messfehlertheorie und ermöglicht über ihre Definitionen die Bestimmung der Reliabilität eines Tests (Bühner, 2011, S. 57). Sie macht keine Aussagen über den Zusammenhang von Itemantworten und Item- und Personenmerkmalen und stellt daher keine Testtheorie im eigentlichen Sinn dar. Ein großes Plus dieser Theorie ist ihre einfache, ökonomische Anwendbarkeit und ihre Bewährtheit. Ihre Grenzen hat sie jedoch bei der Beurteilung der Skalierbarkeit und Konstruktvalidität (Moosbrugger, 2008b).

Weiter ist kritisch anzumerken, dass die Fehlertheorie bzw. die Axiome der KTT nicht psychologisch fundiert sind, d. h. sie wurden nicht aus psychologischen Theorien abgeleitet (Fisseni, 2004, S. 81). Trotz mathematisch korrekter Formulierung sind die Axiome daher in der psychologischen Praxis nicht zwangsläufig haltbar, und zudem lassen sie sich nicht oder nur schwer prüfen (Bühner, 2011, S. 53f.). Die Annahmen der Unkorreliertheit von wahrem Wert und Messfehler, der nur durch zufällige Einflüsse überlagerten Invarianz des wahren Wertes und der Unkorreliertheit der Fehlerwerte über mehrere Messwiederholungen sind in der Realität anzuzweifeln (Bühner, 2011; Embretson & Reise, 2000). Systematische Fehler finden in der KTT keine Berücksichtigung, obwohl sie bekannterweise in der Realität auftreten können, was in solchen Fällen zur Ungültigkeit des Modells der KTT führt und zu einer verfälschten Reliabilitätsschätzung (Bühner, 2011, S. 53f.). Die Annahme stabiler wahrer Werte, beschränkt die Anwendung der KTT zudem auf stabile Merkmale. Die KTT geht darüber hinaus – ohne dies zu prüfen – davon aus, dass die Items eines Tests tatsächlich lediglich eine einzige latente Variable messen und dass z. B. die Summenbildung über die Itemwerte hinweg eine gültige Verrechnungsvorschrift ist (Bühner, 2011, S. 41, 53). Anstatt auf die Eindimensionalität des interessierenden Konstrukts stützt sich die KTT auf die schwächere Annahme der lokalen Unabhängigkeit, was für die Reliabilitätsbestimmung ausreicht. Die Eindimensionalität sollte aber trotzdem nachgewiesen werden.

Zudem ist die Stichprobenabhängigkeit der im Rahmen der KTT berechneten Item-, Test- und Personenstatistiken problematisch, da ihre Generalisierbarkeit schwierig einzuschätzen ist (Embretson & Reise, 2000; Fisseni, 2004). Die Bestimmung der Reliabilität basiert vor allem auf Mittelwerten, Varianzen und Kovarianzen, weshalb eigentlich für jede Stichprobe eine erneute Bestimmung der Reliabilität erforderlich ist (vgl. Bühner, 2011, S. 54). Laut Fisseni (2004, S. 81) bleiben populationsabhängige Aussagen sinnvoll, wobei sie in ihrem Umfang regional und epochal eingeschränkt sind. Durch die Verwendung einer bezüglich des zu erfassenden Merkmals besonders heterogenen Gruppe kann die Reliabilität künstlich erhöht werden, durch die Verwendung einer Gruppe, die bezüglich des zu messenden Merkmals besonders homogen ist, wird sie künstlich gesenkt (Fisseni, 2004, S. 81). Außerdem ist die Reliabilität abhängig von der Testlänge. Das postulierte Skalenniveau (Intervallniveau, da Mittelwerte, Varianzen und Messwertdifferenzen berechnet werden) ist in der Realität oft fragwürdig, und die normbezogene Ergebnisinterpretation ist inhaltlich wenig aussa-

gekräftig (Embretson & Reise, 2000, S. 25ff., 28ff.). Darüber hinaus sind die Folgerungen aus der KTT kritisch zu sehen, da dem sogenannten „Verdünnungsparadox“ zufolge die kriterienbezogene Validität mit steigender Reliabilität von Kriterium und validiertem Test abnimmt (Fisseni, 2004, S. 81).

Trotz allem hat sich die KTT in der Praxis bewährt. Nach Bühner (2011, S. 54) zeichnet sich ein brauchbarer auf der KTT basierender Test in erster Linie durch eine inhaltlich gut begründete Item- und Skalenkonstruktion aus, die die Schwächen der testtheoretischen Annahmen überdecken kann.

IRT. IRT-skalierte Tests bringen einige sehr vorteilhafte Eigenschaften mit. Sie basieren zum einen auf sehr viel strengeren Annahmen als die KTT. Beispielsweise wird die Annahme der lokalen Unabhängigkeit der Items im Rahmen der IRT überprüft, während sie in der KTT mit der Annahme unkorrelierter Fehler für die Berechnung der Reliabilität und Validität einfach vorausgesetzt wird (Bühner, 2011, S. 33). Zum anderen erlauben IRT-skalierte Tests eine konkrete Verhaltensvorhersage, da ermittelt werden kann, mit welcher Wahrscheinlichkeit eine Person ein Item löst, sofern man Itemschwierigkeit und Personenfähigkeit kennt.

Im Rahmen der Rasch-Skalierung wird zudem sichergestellt, dass es sich bei der Summation der Items tatsächlich um eine gültige Verrechnungsvorschrift handelt, und es wird geprüft, ob sich aus dem Summenwert ein hinreichendes Maß der Personenfähigkeit ergibt, was faire Messungen sicherstellt (K. D. Kubinger, 2000). Weitere vorteilhafte Eigenschaften Rasch-skalierter Tests liegen z. B. darin, dass mit ihrer Hilfe Testergebnisse kriteriumsorientiert interpretiert werden können, und darüber hinaus können aufgrund der Berücksichtigung des größeren Messfehlers in den Extrembereichen für die Personenparameter präzisere Konfidenzintervalle angegeben werden als für die Normwerte der KTT.

Umstritten ist jedoch, ob die Testergebnisse von IRT-Skalen tatsächlich ein höheres Skalenniveau als Ordinalwerte aufweisen (Fisseni, 2004, S. 117). Auch stellen sich die Fragen, ob nicht möglicherweise bei der strengen Item- und Personenselektion, die im Rahmen der IRT-Skalierung vorgenommen wird, nicht eigentlich ein „neues“ Testmerkmal konstruiert wird, und, ob eine höhere psychometrische Qualität auch zwangsläufig für diagnostische Fragestellungen adäquater ist.

Unabhängig von der zugrunde gelegten Testtheorie ist zu berücksichtigen, dass die Testkonstruktion nie ganz objektiv stattfindet. Schon die Auswahl einer Definition des zu messenden Konstrukts aus häufig mehreren zur Verfügung stehenden Definitionen sowie die zur Erfassung dieses Konstrukts gewählte Möglichkeit der Operationalisierung sind subjektiv beeinflusst. Hinzu kommt, dass es keine klaren Kriterien zur Itemselektion gibt und die Gewichtung von statistischer Itemgüte und inhaltlicher Bedeutsamkeit für die umfassende Abbildung des Konstrukts ebenfalls sehr willkürlich stattfindet. Thissen und Orlando (2001, S. 90f.) diskutieren in diesem Zusammenhang die

generelle Frage, ob ein Test so konstruiert werden sollte, dass er modellkonform ist, oder ob Testitems inhaltlich ausgewählt werden sollten und anschließend ein Modell zu suchen ist, das zu den Daten passt. Dieser Unterschied in der Herangehensweise stellt auch den zuvor erwähnten Unterschied zwischen dem Rasch-Modell und dem 1-PL-IRT-Modell dar. Weiter sollten Tests unabhängig von der Testtheorie die klassischen Testgütekriterien erfüllen (vgl. z. B. Moosbrugger, 2008a). Gängige Testgütekriterien und Möglichkeiten ihrer Überprüfung sowohl im Rahmen der KTT als auch im Rahmen der IRT werden in den folgenden Abschnitten behandelt.

5.3.2 Testgütekriterien

Es existieren allgemein anerkannte Kriterien, die eine Beurteilung der Testgüte, die Auswahl eines geeigneten Tests sowie eine angemessene Interpretation der Testergebnisse ermöglichen (vgl. Bühner, 2011, S. 58). Diese Gütekriterien werden an dieser Stelle ausführlich beschrieben, da sie die Grundlage für die nachfolgende kritische Betrachtung verfügbarer Lesetests sowie für die im zweiten Teil der Arbeit beschriebene Testkonstruktion bilden. Dabei wird im Folgenden – wie üblich – zwischen Haupt- und Nebengütekriterien unterschieden.

5.3.2.1 Hauptgütekriterien

Traditionell werden drei Gütekriterien angeführt: Objektivität, Reliabilität und Validität (J. Rost, 2004, S. 33). Sie werden meist als Hauptgütekriterien bezeichnet. Bühner (2011) zählt auch die Skalierbarkeit aufgrund ihrer großen Bedeutung zu den Hauptgütekriterien. Dies wird auch im Folgenden so gehandhabt.

Objektivität. Die Objektivität bezieht sich auf den Grad der Unabhängigkeit eines Testergebnisses von verschiedensten Einflüssen außerhalb der zu testenden Person, also z. B. vom Testleiter, der auswertenden Person oder den situativen Bedingungen (J. Rost, 2004, S. 33). Die Objektivität ist eine logische Voraussetzung für die Zuverlässigkeit und Genauigkeit eines Tests. Hängt das Testergebnis vom Testleiter, von situativen Bedingungen, der auswertenden Person etc. ab, kann der Test auch nicht zuverlässig und genau messen (J. Rost, 2004, S. 39). In der Regel werden drei Arten der Objektivität unterschieden: Die Unabhängigkeit des Testergebnisses von der durchführenden Person (Durchführungsobjektivität), von der auswertenden Person (Auswertungsobjektivität) und von der interpretierenden Person (Interpretationsobjektivität) (J. Rost, 2004, S. 39).

Die Testdurchführung sollte zur Sicherung der Objektivität standardisiert ablaufen, was eine genaue Beschreibung der Durchführungsbedingungen, des Auswertungsprozederes und der Interpretationsmöglichkeiten erfordert (Bühner, 2011, S. 58). Hierfür sind unter anderem ein ausführliches und verständliches Testmanual mit kla-

ren Anweisungen und wörtlich wiederzugebenden Instruktionen erforderlich sowie eine leichte Handhabbarkeit der Testmaterialien.

Reliabilität. Reliabilität bedeutet Zuverlässigkeit und meint das Ausmaß an Genauigkeit, mit dem ein Test misst, unabhängig davon, ob es sich dabei um das Merkmal handelt, das er messen soll (J. Rost, 2004, S. 33). Es geht dabei rein um die Messgenauigkeit im Sinne der numerischen Präzision, also um die Zuverlässigkeit, mit der bei einer Messwiederholung unter exakt gleichen Bedingungen dasselbe Ergebnis erzielt wird. Der Wert für die Reliabilität liegt zwischen 0 und 1, wobei ein Wert von 1 für eine perfekt genaue Messung steht und ein Wert von 0 bedeutet, dass keine Messung im eigentlichen Sinn stattgefunden hat (Bühner, 2011, S. 51). Die Reliabilität ist damit eine Voraussetzung für das Kriterium der Validität, auf das anschließend noch eingegangen wird (J. Rost, 2004, S. 33). Ein Test, der ein Kriterium sehr unzuverlässig misst, kann z. B. keine gute Vorhersage leisten.

Im Falle quantitativer und mindestens intervallskalierter Personenvariablen kann die Reliabilität als das Verhältnis von wahrer Varianz zu beobachteter Varianz betrachtet werden (J. Rost, 2004, S. 38). Dabei ist die wahre Varianz die Varianz der nicht beobachtbaren wahren (messfehlerfreien) Testergebnisse, und die beobachtete Varianz ist die Varianz der tatsächlich gemessenen Testergebnisse. Der Wert der Reliabilität gibt damit an, welcher Anteil an der Messwertvarianz tatsächlich auf Unterschiede der Personen zurückzuführen ist (J. Rost, 2004, S. 39). Somit kann derselbe Fehlervarianzbetrag bei einer geringen Streuung der Messwerte relativ groß sein, während er bei einer großen Streuung der Messwerte relativ klein ist (J. Rost, 2004, S. 376). Da die Streuung der Messwerte von der Varianz des Merkmals in einer Population abhängt, ist die Testreliabilität populationsabhängig. In Populationen mit einer größeren Varianz des interessierenden Merkmals fällt die Reliabilität höher aus. Klassischerweise werden zur Schätzung der Reliabilität eines Tests folgende Methoden unterschieden: Verschiedene Konsistenzmethoden, die Retestmethode und die Paralleltestmethode (z. B. Bühner, 2011; J. Rost, 2004). Im Rahmen der IRT stehen zusätzlich weitere Methoden zur Verfügung. Davon sollen im Folgenden die Erwartungswert-Methode und die Reliabilitätsschätzung mithilfe der expected a posteriori- (EAP-) Methode vorgestellt werden (J. Rost, 2004, S. 380ff.), da diese auch in der vorliegenden Arbeit verwendet wurden.

Die *Konsistenzmethoden* beziehen sich auf die gegenseitige Korrelation bzw. Kovarianz von Items oder Testteilen (Bühner, 2011, S. 157). Sie erlauben einerseits eine Aussage über die Qualität der Tests unabhängig von äußeren Bedingungen, und andererseits darüber, innerhalb welcher Grenzen der Messfehler eines Testergebnisses liegt – sofern andere Fehler, z. B. durch Motivation nicht einfließen (Lienert & Raatz, 1998, S. 201). Im Folgenden wird auf spezifische Konsistenzmethoden eingegangen: Verschiedene Split-Half-Methoden und die Methode der internen Konsistenz. Bei den sogenannten „Split-Half"-Methoden wird ein Test bei einer Stichprobe durchgeführt

und der Test anschließend so in zwei Hälften aufgeteilt, dass diese als parallele Messungen gelten können (vgl. Bühner, 2011, S. 157). Dann werden die Rohwerte der beiden Testteile korreliert und der daraus resultierende Koeffizient mithilfe einer Korrekturformel ausgewertet, um die verkürzte Testlänge zu berücksichtigen. Dabei gibt es verschiedene Möglichkeiten, einen Test in zwei parallele Hälften zu teilen (vgl. z. B. Bühner, 2011, S. 157): Bei der Odd-Even-Methode bilden die Items mit gerader Reihungsnummer die eine Hälfte, die Items mit ungerader Reihungsnummer die andere. Weitere Möglichkeiten bestehen darin, die Items zufällig den Testhälften zuzuweisen oder Itemzwillinge mit gleicher Schwierigkeit und Trennschärfe zu bilden und jedes Paar auf die Testhälften aufzuteilen. Darüber hinaus können die Items nach Testzeit aufgeteilt werden, wobei z. B. nach der Hälfte der Testzeit ein Zeichen gegeben wird, auf das hin die Probanden markieren, wie weit sie bis dahin gekommen waren, und anschließend weiter arbeiten. Welche Teilungsmethode die beste ist, hängt von den Eigenschaften des Tests ab. Eine Aufteilung nach Zeit macht z. B. bei Speedtests wenig Sinn.

Die Methode der internen Konsistenz ist die differenzierteste Methode, um die Konsistenz eines Tests zu überprüfen, und weniger willkürlich als die Aufteilung des Tests in zwei Hälften (J. Rost, 2004, S. 379). Meist wird als Index der sogenannte „Cronbachs Alpha“-Wert berechnet. Dabei wird der Test in so viele Teile zerlegt, wie er Items enthält und jedes Item als eigenständiger Testteil betrachtet (J. Rost, 2004; Bühner, 2011). Es wird geprüft, wie sehr die Testitems als Messung einer einzelnen latenten Variablen angesehen werden können. Die Berechnung erfolgt über die Itemvarianzen und -kovarianzen (Bühner, 2011, S. 158). Für unterschiedliche Datenniveaus stehen entsprechende Abwandlungen der Formel zur Verfügung, z. B. die Kuder Richardson 20-Formel (KR-20) für dichotome Daten (Lienert & Raatz, 1998, S. 193). Die Bestimmung der internen Konsistenz ist nur sinnvoll, wenn ein homogenes Merkmal bzw. ein eindimensionales Konstrukt erfasst werden soll. Die Höhe des Werts der internen Konsistenz hängt von der Homogenität der Items und der Testlänge ab. Bei Items mit identischen Varianzen und Fehlervarianzen ist Cronbachs Alpha ein adäquater Schätzer der Reliabilität, ansonsten stellt Cronbachs Alpha lediglich die Untergrenze der Skalenreliabilität dar.

Die *Retestmethode* besteht in der wiederholten Durchführung desselben Tests bei derselben Stichprobe und der Korrelation der Ergebnisse der Testzeitpunkte. Dabei ist auf einen angemessenen Zeitabstand zu achten, damit es nicht zu Wiederholungseffekten kommt. Da nicht ausgeschlossen werden kann, dass in der Zwischenzeit Lern- oder Veränderungsprozesse stattfinden, muss nicht zu beiden Messzeitpunkten dieselbe Variable mit derselben Genauigkeit gemessen werden (J. Rost, 2004, S. 378). Somit fließt in die Retestkorrelation neben der Reliabilität des Tests auch die Stabilität des Merkmals ein. Die *Paralleltestmethode* basiert dagegen auf der Idee, dass die Korrelation der Ergebnisse zweier Tests, die dieselbe Variable in derselben Stichprobe mit gleicher Genauigkeit messen, der Reliabilität dieser Tests entspricht (J. Rost, 2004, S.

377). Das Problem dieser Methode besteht darin, über zwei parallele Tests verfügen zu müssen. Hinzu kommt, dass die Tests von derselben Stichprobe durchgeführt werden müssen, was zur Beeinträchtigung der Reliabilitätsbestimmung durch z. B. Ermüdung, Motivationsabfall oder Übungseffekte führt.

Im Rahmen der IRT können die Fehlervarianz und/oder die wahre Varianz direkt geschätzt werden; ein Umweg über Korrelationen – wie bei der KTT – ist nicht erforderlich (J. Rost, 2004, S. 380). Bei der *Erwartungswertmethode* wird die Fehlervarianz über den Erwartungswert der Standardschätzfehler der Personenparameter und somit unabhängig von der Varianz der beobachteten Messwerte bestimmt (J. Rost, 2004, S. 380). Die Fehlervarianz der einzelnen Personenmesswerte ist dabei unabhängig von der Stichprobe und basiert alleine auf der Anzahl und der Schwierigkeit der Testitems. Eine weitere IRT-Reliabilitätsschätzung basiert auf einer Parameterschätzung nach der *EAP/PV-Methode* (J. Rost, 2004, S. 382). EAP steht für „expected a posteriori", PV für „plausible value". Die Schätzung erfolgt über das Verhältnis der EAP-Streuung zur PV-Streuung. Als EAP-Schätzer bezeichnet man dabei die auf Basis der latenten Verteilung ermittelten optimalen Schätzwerte für die individuellen Messwerte.

Es gibt keine allgemein gültigen Regeln, wie hoch Reliabilitätswerte sein müssen, damit ein Test als reliabel gelten kann. Laut Lienert und Raatz (1998, S. 269) ist ein Reliabilitätskoeffizient von $r_{tt} = .70$ für die Beurteilung interindividueller Differenzen gerade noch ausreichend. Standardisierte Tests sollten eine Konsistenz von $r_{tt} \geq .90$ und einen Parallel- oder Retestwert von $r_{tt} \geq .80$ aufweisen. Für die Beurteilung von Gruppendifferenzen wird eine Reliabilität von $r_{tt} \geq .50$ als akzeptabel angesehen. Laut Bühner (2011, S. 80f.) sind Reliabilitätskennwerte zwischen $r_{tt} = .70$ und $r_{tt} = .80$ niedrig, aber noch ausreichend, Werte zwischen $r_{tt} = .80$ und $r_{tt} = .90$ mittelhoch und Werte von $r_{tt} \geq .90$ hoch. Diese Richtlinien werden in der vorliegenden Arbeit zur Beurteilung der Reliabilitätskoeffizienten herangezogen.

Validität. Die Validität oder Gültigkeit bezieht sich auf den Grad, zu welchem der Test das Merkmal misst, das er zu messen vorgibt, also auf die Aussagekraft des Testergebnisses hinsichtlich der Messintention (J. Rost, 2004; Bühner, 2011). Dabei werden meist drei Validitätsarten unterschieden: Inhaltsvalidität, Konstruktvalidität und Kriteriumsvalidität.

Ein Test ist *inhaltlich valide*, wenn seine Items das zu messende Merkmal hinreichend genau erfassen, also das gesamte Konstrukt (aber kein anderes oder weiteres Konstrukt) abbilden (Bühner, 2011, S. 61f.). Für die Inhaltsvalidität wird kein Kennwert berechnet; die Bestimmung erfolgt aufgrund fachlich-logischer Überlegungen. Augenscheinvalidität, also dass sich auch einem Laien der Zusammenhang zwischen den Testaufgaben und der Messintention unmittelbar erschließt, reicht als Beleg für inhaltliche Validität nicht aus (Bühner, 2011, S. 62). Bei der *Konstruktvalidität* werden häufig Korrelationen mit Ergebnissen anderer Tests und Faktorenanalysen herangezogen (Bühner, 2011, S. 63f.). Dabei werden in der Regel konkrete Erwartungen

über den Zusammenhang des Tests mit konstruktnahen und konstruktfernen Tests formuliert und überprüft. Im Sinne konvergenter Validität sollten die Testergebnisse mit den Ergebnissen von konstruktnahen Tests hoch korrelieren, während sie mit den Ergebnissen konstruktferner Tests im Sinne diskriminanter Validität niedrig oder gar nicht korrelieren sollten. Die faktorielle Validität soll homogene konstruktnahe Inhaltsbereiche zusammenfassen und konstruktferne von diesen trennen (Bühner, 2011, S. 64). Problematisch ist bei der Bestimmung der Konstruktvalidität, dass diese auf das Vorhandensein einer entsprechenden Güte der herangezogenen Tests angewiesen ist, welche nicht zwangsläufig gegeben ist. Als *Kriteriumsvalidität* bezeichnet man den Zusammenhang des Tests mit einem oder mehreren Außenkriterien (häufig z. B. Schulnoten), mit denen aufgrund der Messintention ein hoher Zusammenhang erwartet wird (Bühner, 2011, S. 63).

Bezüglich der Validität gibt es ebenfalls keine starren Normen, ab wann die Werte als ausreichend oder gut bezeichnet werden können; Lienert und Raatz (1998, S. 269ff.) geben jedoch einige Richtlinien als Orientierungshilfe an. Demnach sollte ein Test aus statistischer Sicht, um eine gute Vorhersage gewährleisten zu können, Validitätskoeffizienten von $r_{tc} \geq .70$ aufweisen, wohingegen in der Praxis Werte von $r_{tc} \geq .60$ bereits als sehr zufriedenstellend gelten müssen. Generell ist zu berücksichtigen, für welchen Zweck der Test eingesetzt wird und welche Bedeutung seinem Ergebnis zukommt. Für eine Individualbegutachtung sowie bei einem hohen Gewicht des Testergebnisses für eine weitreichende Entscheidung ist eine wesentlich höhere Validität notwendig als z. B. für eine nicht-individuelle Auslese oder Gruppenvergleiche. Geht es nicht um eine Prognose, sondern eine Diagnose, ist die inhaltliche Validität wichtiger als die praktische, wobei für erstere kein Validitätskoeffizient ermittelt werden kann (Lienert & Raatz, 1998, S. 271). Fisseni (2004, S. 80) bewertet Validitätswerte von $r_{tc} < .40$ als niedrig, Werte von $r_{tc} = .40$ bis $r_{tc} = .60$ als mittelhoch sowie Wert von $r_{tc} > .60$ also hoch. Diese Werte werden auch in der vorliegenden Arbeit als Interpretationshilfe genutzt.

Skalierbarkeit. Skalierbarkeit bedeutet, dass das Testergebnis aufgrund einer gültigen Verrechnungsvorschrift gebildet wird (Bühner, 2011, S. 67). Entsprechend ist zu prüfen, ob z. B. das Aufsummieren der Itemwerte zu einem Gesamtwert legitim ist in dem Sinne, dass die Summenwerte bzw. Unterschiede der Summenwerte tatsächlich die empirisch vorliegenden intra- und interindividuellen Leistungs- oder Merkmalsunterschiede abbilden.

Bei Nominalskalenniveau der Daten ist somit keine Skalierung möglich, da kein Vergleich der Personen im Sinne von besseren oder schlechteren Leistungen erfolgen kann. Damit eine leistungsfähigere Person ein besseres Testergebnis erhalten kann, muss die Messung mindestens auf Ordinalskalenniveau erfolgen. Erfolgt die Messung auf Intervallskalenniveau, ist es zudem möglich, das Ausmaß inter- und intraindividueller Differenzen zu beurteilen. Relationen zwischen Testleistungen können nur bei

Messungen auf Rationalskalenniveau (z. B. Reaktionszeiten) bestimmt werden. Somit hängt die Umsetzbarkeit des Gütekriteriums der Skalierbarkeit vom Skalenniveau des Tests ab (Moosbrugger & Kelava, 2008, S. 18f.). Darüber hinaus spielt die Trennschärfe der Items für die Skalierbarkeit eine wichtige Rolle (Bühner, 2011, S. 68). Die Trennschärfe gibt Auskunft über den Beitrag, den ein Item zum Gesamtskalenwert leistet. Sind die Trennschärfen sehr unterschiedlich, muss gegebenenfalls bei der Summenbildung eine entsprechende Gewichtung der Itemwerte vorgenommen werden. Weiter ist die Problematik der Ratewahrscheinlichkeit unter dem Gesichtspunkt der Skalierbarkeit zu betrachten. Erreichen Personen mit Raten das gleiche Testergebnis wie Personen, die aufgrund von Wissen oder Können die entsprechenden Items richtig gelöst haben, spiegeln gleiche Ergebniswerte nicht gleiche Leistungen wider.

Die Skalierbarkeit kann im Rahmen der IRT mithilfe von Modelltests überprüft werden. Dabei wird untersucht, ob das Verhalten der Testpersonen einem bestimmten mathematischen Modell folgt (Moosbrugger & Kelava, 2008, S. 19). Ist das Kriterium der Skalierbarkeit nicht erfüllt, ist keine valide Messung möglich.

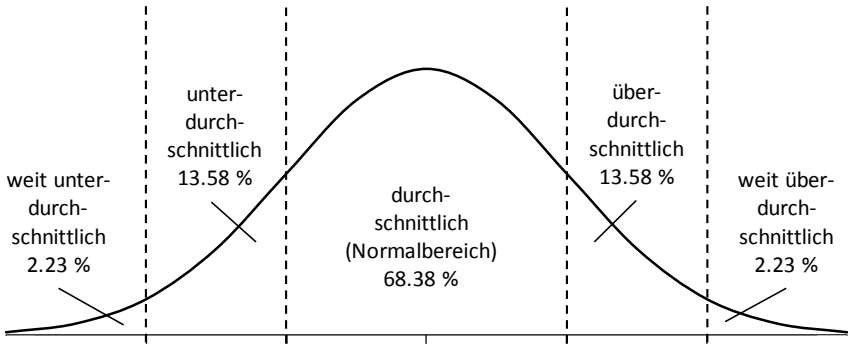
5.3.2.2 Nebengütekriterien

Als Nebengütekriterien werden in der Regel folgende nicht ganz so elementare, aber doch zu berücksichtigende Aspekte genannt: Normierung, Vergleichbarkeit, Ökonomie, Nützlichkeit, Zumutbarkeit, Fairness und Nicht-Verfälschbarkeit.

Normierung. Referenzwerte einer Normstichprobe ermöglichen eine angemessene und objektive, von einem subjektiv gesetzten Kriterium unabhängige Interpretation von Testergebnissen und erlauben eine Aussage über die relative Leistung einzelner Personen (J. Rost, 2004, S. 41). Als Referenzgruppe dient häufig z. B. die Gruppe der Gleichaltrigen oder der Schüler einer Klassenstufe und/oder einer Schulart und/oder gleichen Geschlechts. Mithilfe einer angemessenen Testnormierung wird es möglich, eine Aussage darüber zu treffen, ob eine Person in einem Test unterdurchschnittlich, durchschnittlich oder überdurchschnittlich im Vergleich zur Referenzgruppe abgeschnitten hat (Bühner, 2011, S. 71).

Um das Testergebnis einer Testperson besser interpretieren zu können, werden die Rohwerte in Normwerte umgerechnet. Geläufige Normwertskalen sind z. B. die z-Skala, die T-Wertskala und Prozentränge. Die z-Skala drückt Abweichungen vom Mittelwert in Standardabweichungen aus und hat entsprechend einen Mittelwert von 0 und eine Standardabweichung von 1. Die T-Wertskala hat einen Mittelwert von 50 und eine Standardabweichung von 10. Prozentränge geben an, wie viel Prozent der Personen aus der Normstichprobe genauso gute Leistungen wie oder schlechtere Leistungen als die Testperson erbracht haben. Prozentränge sind besonders leicht zu interpretieren, sie sind jedoch nicht intervallskaliert und erlauben daher keine Bildung von Mittelwerten oder Differenzen. Dafür setzen sie keine Normalverteilung voraus

und können auch für schiefverteilte Messwerte berechnet werden (Bühner, 2011, S. 264). Die verschiedenen Normwerte lassen sich leicht ineinander transformieren (s. Abb. 4).



z-Skala	-2	-1	0	1	2
T-Wertskala	30	40	50	60	70
Prozentrang (gerundet)	2	16	50	84	98

Abbildung 4. Flächenanteile unter der Normalverteilung und Zuordnung der Normwerte sowie Beurteilung nach europäischer Norm.

Ab wann ein Wert als außergewöhnlich angesehen wird, ist eine Frage der vorherrschenden Konventionen. In Europa wird üblicherweise der Bereich von -1 bis 1 Standardabweichung um den Mittelwert als Normalbereich bezeichnet. Dies entspricht einem z-Wertbereich von -1 bis 1, einem Prozentrangbereich von 16 bis 84 und einem T-Wertbereich von 40 bis 60. Da kein Test absolut genau misst, sollten für Normwerte sogenannte „Konfidenzintervalle“ angegeben werden. Diese dienen der Berücksichtigung des Messfehlers. Die Intervalle geben die Bereiche an, innerhalb derer die wahren Leistungen eines Schülers mit einer bestimmten Sicherheit liegen. Je niedriger die Sicherheit gesetzt wird, desto kleiner ist der Bereich. Je höher die Sicherheit gewählt wird, desto größer wird das Konfidenzintervall. Konfidenzintervalle für T-Werte werden auch als T-Wertbänder bezeichnet. Für Prozentränge können keine Konfidenzintervalle berechnet werden (Bühner, 2011, S. 265).

K. D. Kubinger (2006, S. 64) nennt drei Kriterien für die Bewertung einer Testnormierung: Aktualität der Normen, Definition der Population sowie Repräsentativität der Stichprobe. In Bezug auf die Aktualität ist zu sagen, dass Normen von Lesetests nicht älter als zehn Jahre sein sollten, da ältere Lese- und Rechtschreibtests die Leistung der Schüler unterschätzen (Strehlow & Haffner, 2002; Deimel, 2002). Bühner (2011, S. 72) empfiehlt bei Leistungstests sogar generell alle acht Jahre eine Überprüfung der Testnormen vorzunehmen. Bezüglich der Definition der Population ist wichtig, ge-

nau zu beschreiben, für wen die Normen gelten sollen bzw. können (K. D. Kubinger, 2006, S. 64). Meist werden nur Hinweise in Bezug auf notwendige Stichprobengrößen gegeben. Bühner (2011, S. 72) nennt z. B. als Faustregel für die Mindestgröße einer Normstichprobe 300 Personen, Fisseni (2004, S. 80) gibt als generelle Richtlinie zur Beurteilung der Stichprobengröße an, dass Normstichproben von $N < 150$ als klein, Normstichproben von $N = 150$ bis $N = 300$ als mittelgroß und Normstichproben von $N > 300$ als groß bezeichnet werden können. Bühner (2011, S. 72) empfiehlt weiter, neben der Größe der Stichprobe auch die Zusammensetzung der Stichprobe im Manual zu beschreiben, damit der Anwender entscheiden kann, ob die vorliegenden Normen für die von ihm zu testende(n) Person(en) geeignet sind. Vom Testkuratorium der Föderation Deutscher Psychologenvereinigungen wird zudem empfohlen, über die Zielpopulation hinaus auch die Stichprobenziehung genau zu beschreiben (Reimann, 2009, S. 91, 102). Bühner (2011, S. 72) weist zudem darauf hin, dass Normen für alle interessierenden Personengruppen vorliegen sollten.

Die Normierung steht in keinem Zusammenhang mit den Hauptgütekriterien und macht keine Aussage darüber, ob der Test auch etwas Sinnvolles misst (J. Rost, 2004, S. 42). Zudem wird kritisch diskutiert, ob es überhaupt tatsächlich repräsentative Normstichproben gibt, da sich z. B. Testverweigerer in der Regel prinzipiell von den Teilnehmern unterscheiden (vgl. Wyschkon, 2011, S. 321f.). Eine Normierung ist auch nicht für alle Tests notwendig. Bei sogenannten „kriterienorientierten Tests“ kann das Testergebnis mit einem vorab festgelegten, inhaltlich definierten Kriterium verglichen werden (Bühner, 2011; J. Rost, 2004). In manchen Fällen ist dies sinnvoller als ein Vergleich mit einer Referenzgruppe. Für Fragestellungen, die sich auf Mittelwertsdifferenzen zwischen Gruppen beziehen oder auf die Untersuchung von Zusammenhängen, ist ebenfalls keine Normierung nötig.

Um nicht nur auf soziale Vergleichsnormen angewiesen zu sein und Unterschiede in der quantitativen Ausprägung besser greifbar und die Kennwerte inhaltlich interpretierbar zu machen, werden bei IRT-skalierten Tests teilweise sogenannte „Niveaustufen“ definiert (vgl. D. H. Rost & Buch, 2010; W. Lenhard, 2013). Diese beschreiben charakteristische Anforderungen von Aufgaben eines bestimmten Schwierigkeitsgrades. Schüler, die ein bestimmtes Niveau erreichen, lösen also maximal Aufgaben des entsprechenden Schwierigkeitsgrades mit den beschriebenen Anforderungen, wobei alle Anforderungen der darunter liegenden Niveaus eingeschlossen sind. Niveaustufen bieten den Vorteil, kriteriumsorientiert und stichprobenunabhängig Aussagen über das Leistungsniveau einzelner Schüler machen zu können, was aus didaktischer Sicht wünschenswert ist (Baumert et al., 2001). Zudem erlauben sie es, Vergleiche bezüglich der Leistungsverteilung in verschiedenen Ländern anzustellen und Ergebnisse von Schulleistungsstudien für die Öffentlichkeit verständlicher darzustellen (W. Lenhard, 2013, S. 77, 79). Vor allem die groß angelegten internationalen Vergleichsstudien griffen diesen Ansatz auf, was am Beispiel des PISA-Lesetests später noch deutlich wird.

Für derartige Erhebungen sind Niveaustufen auch gedacht; für die Leseverständnisdiagnostik auf Individualebene ist die Gültigkeit der Interpretation von Niveaustufen dagegen kritisch zu sehen, wie W. Lenhard (2013, S. 76ff.) ausführlich darstellt. So resultiert die Zuordnung der Testitems zu den Niveaustufen aus Experteneinschätzungen, wobei die Experten auch zu sehr unterschiedlichen Ergebnissen kommen können. Für ein gutes Niveaustufenmodell ist jedoch eine hohe Übereinstimmung erforderlich. Mangelnde Expertenübereinstimmung kann z. B. darin begründet liegen, dass sich die Items nicht eindeutig einer Niveaustufe zuordnen lassen. Häufig hängt die Schwierigkeit eines Items zudem von multiplen Faktoren ab (z. B. Textgenre, Textgestaltung, Vorwissen). Stimmen die Experten nicht im gewünschten Maß überein oder entspricht die in der Normierung ermittelte Rangfolge der Aufgabenschwierigkeiten nicht der vom Modell postulierten, werden häufig einfach die unpassenden Items aus dem Test entfernt, unabhängig von ihren psychometrischen Eigenschaften oder ihrer inhaltlichen Bedeutung für die Abbildung des zu erfassenden Konstrukts. Darüber hinaus werden zum Teil großzügige Überlappungsbereiche zugelassen, und die Veranschaulichung komplexer psychometrischer Kennwerte zu einer Niveaustufe geht zulasten der Messgenauigkeit und Differenzierungsfähigkeit. Zudem bleibt der Messfehler unberücksichtigt, da keine Konfidenzintervalle für die Niveaustufen bestimmt werden. Eine Vergleichbarkeit der Niveaustufen verschiedener Tests ist aufgrund unterschiedlicher zugrunde liegender Modelle in der Regel nicht gegeben. Bei IRT-skalierten Tests trifft es zudem in der Realität nicht immer, sondern nur mit einer bestimmten Wahrscheinlichkeit zu, dass tatsächlich alle unter dem erreichten Niveau liegenden Aufgaben korrekt gelöst wurden. Die einzelnen Testaufgaben werden bei der Bestimmung der erreichten Niveaustufe jedoch nicht berücksichtigt, was bei der Beurteilung der Testergebnisse auf Individualebene problematisch sein kann. Weiter merken D. H. Rost und Buch (2010) an, dass Niveaustufen den Anschein erwecken, es handle sich um diskrete, qualitativ unterschiedliche Stufen, was in Bezug auf das Leseverständnis dem aktuellen Forschungsstand widerspricht. Es handelt sich also um eine künstliche Stufenbildung in einem Kontinuum (Turner, 2002).

Manchmal werden jedoch bei IRT-skalierten Tests nicht (nur) Niveaustufen, sondern (auch) *Personenparameter* als Vergleichswerte in Testmanualen angegeben. Der Personenparameter gibt für ein bestimmtes Testergebnis die Summe der Lösungswahrscheinlichkeiten über alle Items unter Geltung des Rasch-Modells an (Bühner, 2011, S. 501). Je höher die Personenfähigkeit im Vergleich zur Schwierigkeit eines einzelnen Items ausfällt, desto größer ist die Wahrscheinlichkeit, dass eine Person das Item richtig löst. Da sich die Logit-Lösungswahrscheinlichkeit für ein Item aus der Differenz von Item- und Personenparameter ergibt und die durchschnittliche Personenfähigkeit auf 0 festgelegt ist, kennzeichnen negative Personenparameter eine geringere Fähigkeit und positive Werte eine höhere Fähigkeit der Person im Vergleich zur Gesamtstichprobe. Die Testergebnisse können somit auch bezogen auf die durchschnittliche Personenfähigkeit interpretiert werden, wobei die Höhe des Betrags das Ausmaß der Ab-

weichung von der durchschnittlichen Schülerleistung angibt und das Vorzeichen die Richtung der Abweichung (vgl. Goldhammer & Hartig, 2008; Moosbrugger, 2008a). Somit ermöglichen Personenparameter nicht nur eine kriteriumsorientierte Ergebnisinterpretation, sondern auch den Vergleich der Schüler hinsichtlich ihrer Leistung. Darüber hinaus kann bei Personenparametern für jeden Messwert ein individuelles und somit präziseres Konfidenzintervall angegeben werden als für die gemäß KTT bestimmbaren Normwerte. Dabei wird berücksichtigt, dass Personenparameter in den Extrembereichen stärker messfehlerbehaftet sind als Personenparameter im mittleren Leistungsbereich (Bühner, 2011, S. 567).

Vergleichbarkeit. Mit dem Gütekriterium der Vergleichbarkeit ist gemeint, dass eine oder mehrere Parallelformen des Tests oder andere Tests mit demselben Gültigkeitsbereich vorliegen, wobei eine Person, die beide Test(-formen) bearbeitet, jeweils ähnliche Ergebnisse erzielen sollte (Bühner, 2011, S. 72). Dieses Kriterium ist häufig nicht erfüllt, da es schwierig ist, tatsächlich parallele Tests zu konstruieren. Zudem ist es fraglich, ob der Aufwand einer zusätzlichen Testkonstruktion unternommen wird, wenn bereits ein Test mit demselben Gültigkeitsbereich vorliegt.

Ökonomie. Als ökonomisch zählt z. B. eine (angemessen) kurze Durchführungsdauer, geringer Materialverbrauch, schnelle und bequeme Auswertbarkeit, die Möglichkeit des Einsatzes als Gruppentest und eine einfache Handhabung (Bühner, 2011, S. 72). Zur Beurteilung der Ökonomie können andere Diagnosemethoden (z. B. Verhaltensbeobachtung und Interview) zum Vergleich herangezogen werden. Der Aufwand sollte immer in einem angemessenen Verhältnis zum Nutzen stehen, zugleich sollte die Ökonomie nie zulasten der Sorgfalt, der sachgerechten Durchführung und der verantwortungsvollen Diagnostik gehen.

Nützlichkeit. Ein Test ist nützlich, wenn er ein praxisrelevantes Kriterium erfasst (Bühner, 2011, S. 73). Steht bereits ein Test zur Erfassung dieses Merkmals oder dieser Verhaltensweise zur Verfügung, ist gut zu begründen, welche Vorteile der neue Test gegenüber dem bereits vorhandenen aufweist.

Zumutbarkeit. Unter Zumutbarkeit wird verstanden, dass die zu testende Person durch die Testung keinen unnötigen physischen, psychischen oder zeitlichen Belastungen ausgesetzt wird (Bühner, 2011, S. 73).

Fairness. Ein Test gilt als fair, wenn die Relation der Testergebnisse die entsprechende Relation des interessierenden Merkmals abbildet und die Testergebnisse nicht zur Diskriminierung einzelner für die Testung relevanter Personen oder Personengruppen führen können (Bühner, 2011, S. 73).

Nicht-Verfälschbarkeit. Für die getesteten Personen sollte es nicht möglich sein, das Testergebnis willentlich oder unbewusst in eine gewünschte Richtung zu verfälschen.

Dies kann zwar nie vollständig ausgeschlossen werden, Items sollten aber z. B. möglichst so formuliert werden, dass sie nicht automatisch sozial erwünschtes Antwortverhalten auslösen (Bühner, 2011, S. 73f.). Wenn möglich, sollte im Testmanual angegeben werden, wie Verfälschungen erkannt werden können. Bei Leistungstests ist in der Regel nur eine Verfälschung nach unten möglich.

5.3.3 Möglichkeiten der Operationalisierung

Nachdem in den obigen Abschnitten testtheoretische Grundlagen und Testgütekriterien erläutert wurden, stellt sich nun die Frage nach der inhaltlichen Gestaltung von Leseverständnistests, d. h. danach, wie die unterschiedlichen Aspekte von Leseverständnis (z. B. Rekodieren, Dekodieren, Leseflüssigkeit, Textverständnis) erfassbar gemacht werden können. Je nach Konzeption und Fragestellung wird Leseverständnis mithilfe von verschiedenartigem schriftsprachlichem Material untersucht. Abhängig davon, ob Pseudowörter, einzelne Wörter, einzelne Sätze oder ganze Texte dargeboten werden, können Rückschlüsse auf unterschiedliche kognitive Prozesse bzw. auf die Dekodierfähigkeit oder Textverarbeitungstiefe gezogen werden. Im Folgenden werden einige Beispiele für Möglichkeiten der Operationalisierung dargestellt.

Rekodieren. Die Rekodierleistung lässt sich anhand von Lautleseaufgaben prüfen. Es können Wörter, Sätze oder Pseudowörter verwendet werden, wobei insbesondere anhand von Pseudowörtern das Beherrschen der Graphem-Phonem-Korrespondenzregeln überprüft werden kann. Je nachdem, ob die Zeit gestoppt und/oder Lesefehler gezählt werden, ist die Erfassung der Rekodiergeschwindigkeit und/oder -genauigkeit möglich. Über die kombinierte Berücksichtigung von Geschwindigkeit und -genauigkeit erhält man eine Einschätzung des Automatisierungsgrades phonologischen Rekodierens (Petermann & Daseking, 2012). Darüber hinaus können beim Lautlesen Fehlertypen identifiziert werden. Inhaltliches Verständnis wird dabei nicht erfasst.

Ein Nachteil der Lautleseaufgaben ist, dass die Lautlesegeschwindigkeit mit der Artikulationsgeschwindigkeit konfundiert ist und sich generell bei Lautleseverfahren eine Vertrautheit mit der Lautlesesituation und das Wohlbefinden in der Lautlesesituation auf die Leistung auswirken können. Zudem sind Lautleseverfahren ausschließlich für Individualdiagnostik geeignet. Lautes Lesen ermöglicht es zwar, neben der Lesegeschwindigkeit auch die Lesegenauigkeit zu erfassen, jedoch ist Letztere lediglich in den ersten Grundschuljahren und insbesondere auch für Sprachen mit inkonsistenter Graphem-Phonem-Korrespondenz zur interindividuellen Differenzierung in der Leseleistung geeignet. Im Gegensatz zur Anzahl der Fehler ist die Anzahl der pro Minute gelesenen Wörter ein guter Indikator für das Leseverständnis (Walter, 2008). Insbesondere auch in der Sekundarstufe hat die Erfassung der Lesegeschwindigkeit mehr Aussagekraft (W. Lenhard, 2013, S. 86).

Dekodieren. Die Dekodierleistung kann über Lautleseaufgaben oder stilles Lesen erfasst werden, wobei jeweils zusätzlich das bedeutungsmäßige Verständnis des Gelesenen zu prüfen ist. Ein Vorteil stiller Leseaufgaben liegt darin, dass diese ökologisch valider sind, da stilles Lesen im Alltag den Regelfall darstellt (W. Lenhard, 2013, S. 88). Außerdem ist bei Tests mit stillen Leseaufgaben ein Einsatz als Gruppentest möglich. Die Dekodieraufgaben können weiter danach differenziert werden, ob die Genauigkeit und/oder die Geschwindigkeit erfasst wird. Bei regulären Orthographien ist vor allem die Dekodiergeschwindigkeit ein guter Prädiktor für die Leseleistung (Florit & Cain, 2011; Landerl & Willburger, 2009).

Auf Wortebene kann die Dekodierleistung überprüft werden, indem z. B. zu einem Wort möglichst schnell aus mehreren Bildern das passende gefunden werden muss. Bei diesem Aufgabentyp spielt jedoch z. B. die Bilderfassung zusätzlich mit hinein. Zudem ist das Lesen von isolierten Wörtern nicht mit dem Lesen von kontextualisierten Wörtern vergleichbar, da gute Leser den Satzkontext zum Wortverständnis hinzuziehen (Klicpera & Gasteiger-Klicpera, 1995, S. 133). Deshalb ist das Lesen isolierter Wörter für die Einschätzung des Leseverständnisses insbesondere in mittleren und höheren Klassenstufen weniger gut geeignet (W. Lenhard, 2013, S. 87).

Leseflüssigkeit. Die Leseflüssigkeit, welche gemäß der Definition von Nix (2011, S. 61) neben der Lesegeschwindigkeit und -genauigkeit auch prosodische Phrasierung und somit ein mindestens lokales Leseverständnis einschließt, kann durch Lautleseaufgaben mit Sätzen oder Texten erfasst werden. Lautleseaufgaben haben dabei den Vorteil, dass zusätzlich erkannt werden kann, ob ein Schüler Fehler bemerkt und selbstständig korrigiert (W. Lenhard, 2013, S. 88). Jedoch besteht dann wieder das Problem, dass Leseflüssigkeitstests nicht als Gruppentest einsetzbar sind und dass Artikulationsgeschwindigkeit sowie die Gewohnheit und das Wohlbefinden beim lauten Lesen eine Rolle spielen. So können Artikulationsprobleme oder auch Schüchternheit trotz gut ausgeprägtem Leseverständnis das expressive Lesen beeinträchtigen. Umgekehrt impliziert die Fähigkeit zum expressiven Lesen nicht zwangsläufig auch das Verständnis des Gelesenen (W. Lenhard, 2013, S. 89).

Lesegeschwindigkeit kombiniert mit Leseverständnis. Eine Möglichkeit zur kombinierten Erfassung von Lesegeschwindigkeit und Leseverständnis ist die Vorgabe eines Lückentextes. Die Lücken befinden sich entweder in festen Abständen oder an Schlüsselstellen des Textes. Dabei ist entweder aus mehreren Optionen das passende Wort auszuwählen oder ein passendes Wort frei zu finden und einzusetzen. Problematisch ist hier nicht nur die Konfundierung von Lesegeschwindigkeit (meist ist eine enge Zeitbegrenzung vorgegeben) und Leseverständnis, sondern auch die hohe Ratewahrscheinlichkeit (meist stehen pro Lücke nur drei Optionen zur Verfügung). Ein derartiger Test ist zwar ökonomisch durchzuführen und auszuwerten und kann als Gruppentest eingesetzt werden, jedoch sind die Korrelationen der einzelnen Komponenten (Anzahl korrekt ausgewählter Wörter bzw. Lesezeit) mit umfassenderen Le-

setestbatterien gering. W. Lenhard (2013, S. 90) empfiehlt daher als zuverlässigeren Indikator den Quotienten „Anzahl korrekter Lösungen durch Zeit“ zu verwenden. Verfahren ohne Vorgabe von Optionen seien wenig aussagekräftig.

Auch im Rahmen von Satzleseaufgaben lassen sich Lesegeschwindigkeit und Leseverständnis kombiniert prüfen. Dabei werden in der Regel nach Schwierigkeit aufsteigend angeordnete Sätze vorgegeben, die anschließend auf ihre inhaltliche Richtigkeit hin beurteilt werden sollen (vgl. z. B. Auer et al., 2005; Mayringer & Wimmer, 2003). Wichtig ist dabei, dass die Sätze sehr leicht als inhaltlich richtig oder falsch beurteilt werden können, um nicht Vorwissen, sondern das Leseverständnis abzufragen. Satzleseaufgaben zum stillen Lesen ermöglichen ebenfalls eine Durchführung in Form von Gruppentests und stellen zudem einen sehr reliablen Indikator für das Leseverständnis dar (W. Lenhard, 2013, S. 90).

Im Gegensatz zu Wortleseaufgaben sind Lückentexte und Satzleseaufgaben auch für höhere Klassenstufen geeignet. Sie haben jedoch den Nachteil, dass basale Lesekompetenzen und Leseverständnis stärker konfundiert erfasst werden, denn um sicherzustellen, dass die Sätze tatsächlich gelesen wurden, ist stets eine Verständnisprüfung notwendig. Sollen hauptsächlich basale Lesekompetenzen erfasst werden, ist es dabei wichtig, die Sätze und Texte einfach zu halten, um nicht zu stark höhere Verarbeitungsprozesse einzubeziehen. Da bei Satzleseaufgaben klar ist, dass die Sätze zusammenhanglos aneinander gereiht sind und der Inhalt der Sätze nicht weiter zur Verfügung stehen muss, sollten hierarchiehöhere Prozesse nicht stattfinden, denn die Bildung globaler Kohärenzen ist nicht erforderlich.

Reines Leseverständnis. Reines Leseverständnis (ohne Lesegeschwindigkeit) lässt sich überprüfen, indem Sätze oder Fließtexte ohne Zeitbegrenzung vorgegeben und Verständnisfragen dazu gestellt werden, oder indem ein Text mündlich oder schriftlich nacherzählt oder zusammengefasst werden soll. Hierbei sind Variationen der Textlänge, des Textgenres und des Textformats ebenso möglich wie des Anforderungsniveaus der Fragen und des Antwortformats. Eine Alternative zur Prüfung der Verarbeitungstiefe sind Satz- oder Textverifikationsaufgaben. Spätestens ab der Sekundarstufe sollten zur Prüfung der Leseleistung (zusätzlich zu Aufgaben zur Prüfung basaler Lesekompetenzen) Aufgaben zur Erfassung höherer Verständnisebenen eingesetzt werden (W. Lenhard, 2013, S. 91).

Werden Verständnisfragen zu einem Text gestellt, können diese sich auf unterschiedliche Ebenen des Textverständnisses beziehen und wörtlich im Text gegebene Informationen oder inhaltlich (z. B. paraphrasiert) im Text enthaltene Informationen erfragen. Verständnisfragen können aber auch so formuliert sein, dass ihre Beantwortung über den Text hinausgehende Schlussfolgerungen bzw. den Rückgriff auf Vorwissen erfordert. Um lokale Kohärenzbildung zu erfassen, genügen mehrere zusammenhängende Sätze bzw. kurze Texte, zu denen Verständnisfragen gestellt werden. Zur Überprüfung globaler Kohärenzbildung sind hingegen längere Texte notwendig,

was wiederum mit einer längeren Testzeit einhergeht. Da unterschiedliche Textgenres unterschiedliche Anforderungen an den Leser stellen, unterschiedliche Erwartungen beim Leser auslösen und sich die Schülerleistungen hinsichtlich der Textgenres unterscheiden, kann es darüber hinaus sinnvoll sein, verschiedene Textgenres einzubeziehen.

Die Fragen können in offenem oder gebundenem Antwortformat dargeboten werden. Beim offenen Antwortformat ist die Antwort von der Testperson selbst zu formulieren, beim gebundenen Antwortformat sind Antwortalternativen vorgegeben, und zutreffende Antworten sind auszuwählen. Ersteres hat das Problem der Konfundierung von Leseverständnis und Schreibfähigkeiten, letzteres das Problem, dass Raten mit einer gewissen Erfolgswahrscheinlichkeit möglich ist. Sogenannte „Multiple-Choice-(MC-) Aufgaben“ führen zudem dazu, dass die Schüler nicht selbst Bedeutung generieren, sondern vorgegebene Bedeutungszuschreibungen auf die Wahrscheinlichkeit ihres Zutreffens hin beurteilen müssen. MC-Aufgaben bilden das Textverständnis also nur indirekt ab (vgl. Spinner, 2004). Ein Vorteil gebundener Antwortformate besteht jedoch darin, dass sie die Beurteilung der Lösung als korrekt oder falsch erleichtern, wenn eindeutig vorgegeben ist, welche Antwortalternativen korrekt oder falsch sind. Dies wirkt sich positiv auf die Objektivität aus. Zudem ist die Auswertung gebundener Aufgaben in der Regel ökonomischer. Nicht zuletzt kann ein Test mit MC-Antwortformat prinzipiell maschinell anhand eines Dokumentenscanners ausgewertet werden. Letzteres ist vor allem für Forschungsprojekte mit sehr großem Stichprobenumfang ein nicht zu verachtender Vorteil. Ist von vornherein klar, dass pro Frage nur eine Antwortalternative korrekt ist, spricht man auch von „Single Choice- (SC-) Format“. Dieses ist den meisten Schülern zwar am geläufigsten (z. B. aus Quizsendungen im Fernsehen), jedoch ist hier die Ratewahrscheinlichkeit höher und es kann prinzipiell nach dem Ausschlussverfahren vorgegangen werden.

Beim mündlichen Nacherzählen oder Zusammenfassen von Gelesenem lässt sich die Bewertung kaum objektivieren, und die Tests sind nur im Individualsetting durchführbar (W. Lenhard, 2013, S. 91). Beim schriftlichen Nacherzählen oder Zusammenfassen ist die Auswertung zwar objektiver möglich, und der Test kann auch als Gruppentest eingesetzt werden, jedoch ist wie bei Fragen mit offenem Antwortformat im Testergebnis das Leseverständnis mit den Schreibfähigkeiten konfundiert. Denn es spielt bei diesen Tests nicht nur das Textverständnis, sondern auch die Fähigkeit eine Rolle, das Verstandene in Worte zu fassen und die eigene Formulierung schriftlich zu fixieren (vgl. Spinner, 2004). Meist ist die Verstehensleistung jedoch besser entwickelt als die Fähigkeit, das Verstandene in Worte zu fassen. Aus diesen Gründen bezeichnet Spinner (2004, S. 134) Schüleräußerungen als „nur ein indirektes und ungenaues Indiz für Verstehensleistungen“.

Bei der Entscheidung für ein schriftliches Antwortformat stellt sich somit die Frage, ob man lieber in Kauf nimmt, Wiedererkennungseffekte zu prüfen (gebundenes Antwortformat), oder ob man eine Konfundierung des Verständnisses mit der Fähigkeit

das Verstandene in Worte zu fassen und zu Papier zu bringen (offenes Antwortformat) präferiert. Sowohl beim offenen als auch beim gebundenen Antwortformat ist zu berücksichtigen, dass die Formulierung und das Verständnis der Frage stets mit einfließen (Spinner, 2004). Somit wäre eigentlich auch die Komplexität der Frageformulierung zu berücksichtigen.

Bei Satzverifikationsaufgaben dagegen kann z. B. durch Verwendung von Satzvariationen (z. B. mithilfe von Synonymen oder Passiv- vs. Aktivkonstruktionen) geprüft werden, ob die Informationen nur auf Oberflächenebene verarbeitet wurden, oder ob eine propositionale Repräsentation aufgebaut wurde. Wird nur die wörtliche Aussage als bereits bekannt empfunden, wurde nur eine Oberflächenrepräsentation aufgebaut. Werden aber auch Variationen in der Formulierung, die den Inhalt nicht verändern, akzeptiert, wird der Text propositional repräsentiert. Auch Sätze, zu deren Verifikation Schlussfolgerungen erforderlich sind, können formuliert werden. Derartige Satzverifikationsaufgaben haben sich als gute Indikatoren für das Leseverständnis erwiesen (Marcotte & Hintze, 2009).

Hörverständnis. Folgt man dem SVR-Modell, müsste man neben dem Dekodieren auch das generelle Sprachverständnis prüfen. Dies wäre z. B. in Form eines Hörverständnistests möglich, der es erfordert, einen gesprochenen Text zu verstehen. Um unabhängig von der Dekodierleistung zu bleiben, müssten die Verständnisfragen ebenfalls mündlich vorgegeben und beantwortet werden, was wiederum die Durchführung als Gruppentest ausschließt. Weiter ist es als Einschränkung zu sehen, dass ein Hörverständnistest zwar unabhängig von der Dekodierleistung ist, jedoch durch das auditive Darbieten des Textes Komponenten hineingebracht werden, die beim Leseverständnis keine Rolle oder eine geringere Rolle spielen (z. B. Funktionsfähigkeit des Gehörs). Darüber hinaus steht beim Lesen der Text in der Regel weiterhin zur Verfügung, während man beim Hören eines Textes sofort nach der Darbietung auf das Gedächtnis angewiesen ist (vgl. Hoover & Gough, 1990).

Abhängig davon, welcher Teilprozess des Lesens von Interesse ist, ist also ein entsprechendes Aufgabenformat zu wählen. Die Auswahl einer entsprechenden Operationalisierungsvariante wirkt sich z. B. auf die Ökonomie eines Tests aus. Soll beispielsweise Leseverständnis umfassend erfasst werden, und sollen sowohl basale Lesekompetenzen als auch höhere Verständnisebenen einbezogen werden, sind verschiedene Aufgabenformate notwendig und die Testzeit fällt relativ hoch aus. Auch ermöglichen manche Operationalisierungen lediglich den Einsatz eines Tests als Individualtest, während andere Varianten auch für Gruppentestungen geeignet sind. Für eine Integration von Lesetests in den Schulalltag sollten Tests ökonomisch und möglichst als Gruppentests einsetzbar sein. Auch für die Verwendung im Rahmen großangelegter Forschungsprojekte ist die Einsetzbarkeit eines Tests als Gruppentest ein großer Vorteil. Individualtests hingegen eigenen sich eher für klinische Diagnostik.

5.3.4 Verfügbare Lesetests für die Sekundarstufe

Nachfolgend wird beschrieben, welche Aspekte von Lesekompetenz in welcher Form von verschiedenen derzeit verfügbaren standardisierten Lesetests erfasst werden, die zumindest für einen Teil der Sekundarstufe geeignet sind. Aufgrund der Vielzahl und ständig neuen Tests wird kein Anspruch auf Vollständigkeit erhoben, sondern lediglich eine Auswahl aktueller, gängiger Tests beschrieben. Dabei werden die wichtigsten Aspekte der Tests sowie die Testgüte betrachtet und kritisch bewertet. Zur Beurteilung der Testgüte werden vor allem Angaben der Autoren zu Merkmalen sowie zu den Gütekriterien in den Testmanualen herangezogen.

Auf die Darstellung des Tests ELFE 1-6 (Ein Leseverständnistest für Erst- bis Sechstklässler; W. Lenhard & Schneider, 2006) wird verzichtet, da dieser Test hauptsächlich für die Grundschule konstruiert wurde und in den Klassenstufen fünf und sechs lediglich als Screening eingesetzt werden kann. Schon in der vierten Klassenstufe kommt es beim Einsatz des Tests zu Deckeneffekten (W. Lenhard, 2013, S. 95).

Ein-Minuten-Leseflüssigkeitstest. Der Ein-Minuten-Leseflüssigkeitstest (Moll & Landerl, 2010; Landerl & Willburger, 2009) ist ein Individualtest für die Klassenstufen eins bis sechs sowie für junge Erwachsene. Der Test ist separat oder als Teil des Salzburger Lese- und Rechtschreibtests II (SLRT-II; Moll & Landerl, 2010) einsetzbar. Der Ein-Minuten-Leseflüssigkeitstest fordert möglichst schnelles und fehlerfreies lautes Lesen einer Wort- bzw. Pseudowortliste innerhalb einer Minute. Die Durchführungsdauer liegt bei maximal fünf Minuten, die Auswertung dauert ebenfalls ca. fünf Minuten. Das Testergebnis besteht in der Anzahl korrekt gelesener Wörter. Es ist möglich, eine separate Diagnose von zwei Teilkomponenten des Wortlesens vorzunehmen: Einerseits Defizite in der automatischen, direkten Worterkennung und andererseits Defizite im synthetischen, lautierenden Lesen.

Der Test liegt in Parallelversionen vor. Die Paralleltestkorrelationen für die Anzahl korrekt gelesener Wörter bzw. Pseudowörter fallen mit Werten zwischen $r = .90$ und $r = .98$ sehr hoch aus. Korrelationen der Testergebnisse mit anderen Lesetests liegen zwischen $r = .69$ und $r = .92$ und sprechen ebenso wie Korrelationen mit der Deutschnote, die zwischen $r = .37$ und $r = .54$ liegen, für eine hohe Validität des Tests. In den Jahren 2006 bis 2008 wurden Normdaten für die erste bis sechste Klasse ($N = 1\,747$) und für junge Erwachsene ($N = 241$) erhoben. Die Daten der fünften und sechsten Klassenstufe stammen aus Haupt- und Realschulen aus Salzburg und Baden-Württemberg.

Beim Ein-Minuten-Leseflüssigkeitstest handelt sich somit um einen sehr reliablen, validen und ökonomischen Individualtest, der bei einer sehr breiten Altersspanne eingesetzt werden kann. Als Individualtest ist er für den Einsatz im Schulalltag jedoch nur eingeschränkt geeignet. Darüber hinaus wird in der fünften und sechsten Klasse nicht das gesamte Leistungsspektrum in die Normierung einbezogen (keine Gymna-

sien), und die Normstichprobe beschränkt sich in diesen Klassenstufen auf lediglich zwei Regionen. Baden-Württemberg macht zudem nur einen kleinen Teil der Stichprobe aus. Somit ist eine Generalisierbarkeit der Ergebnisse auf den gesamten deutschen Sprachraum nicht zwangsläufig möglich. Der Test beschränkt sich zudem auf die Erfassung der Wortlesegeschwindigkeit; das Lesen von Sätzen oder Texten sowie deren Verständnis werden nicht geprüft.

Verlaufsdiagnostikum sinnerfassenden Lesens (VSL). Das VSL (Walter, 2013) ist ein aktueller Test, der der systematischen formativen Evaluation von Unterrichts- und Fördermaßnahmen dienen soll. Er eignet sich laut Autor für Schüler der zweiten bis sechsten Klasse sowie für Förderunterricht, LRS-Förderkurse, LRS-Therapie und Alphabetisierungskurse. Die längsschnittliche Erhebung von Lesekompetenz und die Lernverlaufs- bzw. -fortschrittsdiagnostik sind ebenso möglich wie der Einsatz als Niveautest. Das VSL kann als Screening und insbesondere aufgrund seiner Durchführbarkeit als Gruppentest auch für Forschungszwecke genutzt werden. Die Schüler sollen beim VSL einen Text lesen und für jedes siebte Wort aus einer Klammer mit drei Optionen das in den Kontext passende Wort auswählen. Dafür haben sie vier Minuten Zeit. Zahlreiche Parallelformen ermöglichen einen Einsatz an bis zu 20 Messzeitpunkten. Das VSL ist in Papierform und als PC-Version erhältlich.

Mit Paralleltestkorrelationen von mindestens $r = .77$ und einer internen Konsistenz von $\alpha = .93$ ist der Test als reliabel zu bezeichnen. Im Manual werden zudem positive Befunde zur Änderungssensibilität, zur kriteriumsorientierten Validität, zur kriteriumsorientierten Veränderungsvalidität sowie zur Konstruktvalidität dokumentiert. Für Niveau- und Veränderungsmessungen liegen Prozentränge und T-Werte für die Grund-, Gemeinschafts- und Realschule sowie das Gymnasium vor. Die Normstichprobe besteht aus $N = 3\,036$ (Niveau) bzw. $N = 2\,289$ (Veränderung) Schülern, wobei die nach Klassenstufe und Schulart aufgeteilten Substichproben in der Sekundarstufe z. T. recht klein ausfallen (z. B. Realschule Klasse 5/6: $n = 80$). Aufgrund der kurzen Durchführungsdauer und der leichten Auswertbarkeit ist der Test als sehr ökonomisch anzusehen. Insbesondere bei der PC-Version kann auch der Lernverlauf leicht veranschaulicht werden. Die Nachteile eines derartigen Lückentextes zur Erfassung des Leseverständnisses wurden bereits in Kapitel 5.3.3 besprochen: Lesegeschwindigkeit und -verständnis sind konfundiert, bei nur drei Lösungsoptionen ergibt sich eine hohe Ratewahrscheinlichkeit, und häufig zeigten sich geringe Korrelationen mit umfassenderen Leseverständnistestbatterien (vgl. W. Lenhard, 2013, S. 90).

Frankfurter Leseverständnistest für 5. und 6. Klassen (FLVT 5-6). Mit dem FLVT 5-6 (Souvignier, Trenk-Hinterberger, Adam-Schwebe & Gold, 2008; Adam-Schwebe et al., 2009) lässt sich das Leseverständnis in den Klassenstufen fünf und sechs überprüfen. Er ist als Gruppen- oder Individualtest einsetzbar und besteht aus zwei Teilen: Einem Erzähltext und einem Sachtext von je ca. 560 Wörtern. Zu jedem Text

werden 18 Verständnisfragen im SC-Format gestellt. Die Durchführung des FLVT 5-6 dauert ca. eine Schulstunde.

Der Test liegt in zwei Parallelformen (A/B) vor. Die interne Konsistenz beträgt für Testform A $\alpha = .88$, für Testform B $\alpha = .86$. Der minderungskorrigierte Paralleltestkoeffizient beträgt $r = .80$. Die Validität ist theoretisch begründet, und erwartungskonform fallen Korrelationen mit Ergebnissen anderer Leseverständnistests recht hoch aus ($r = .57$ bis $r = .74$), während Korrelationen mit Ergebnissen eines Mathematiktests erwartungskonform niedrig ausfallen ($r = .26$ bis $r = .27$). Als Normen werden für die fünfte und sechste Klasse der verschiedenen Schularten Prozentränge, z- und T-Werte zur Verfügung gestellt. Die Normstichprobe umfasst für Testform A $N = 1\,239$ Schüler und für Testform B $N = 1\,237$ Schüler. Die Normdaten stammen aus dem Schuljahr 2004/2005. Zusätzlich können die Schüler aufgrund ihres Gesamtergebnisses einer Niveaustufe des Leseverständnisses zugeordnet werden.

Kritisch ist bei diesem Test anzumerken, dass das Manual keine Angabe über die Herkunft der Normstichprobe macht. Auf Anfrage teilte einer der Autoren mit, dass die Normierung ausschließlich in der hessischen Rhein-Main-Region stattfand. Wie sich im Verlauf der vorliegenden Arbeit (z. B. in Kap. 6) zeigen wird, müssen die Ergebnisse eines Bundeslandes nicht zwangsläufig auch auf andere Bundesländer generalisierbar sein. Zudem unterscheiden sich die beiden Testformen leicht in der Schwierigkeit, und basale Lesekompetenzen werden nicht separat berücksichtigt. Darüber hinaus sind die Normen zum Zeitpunkt der Erstellung dieser Arbeit bereits etwa neun Jahre alt und bedürfen somit einer baldigen Erneuerung (vgl. in Kap. 5.3.2.2).

Zürcher Lesetest-II (ZLT-II). Der ZLT-II (Petermann & Daseking, 2012) ist eine Weiterentwicklung des ZLT von Linder und Grissemann (2000) für die Klassenstufen eins bis acht. Er soll der Diagnostik und Verlaufskontrolle von Lesestörungen sowie der Ableitung von Förderempfehlungen dienen. Der Individualtest wurde somit vor allem zur Differenzierung im unteren Leistungsbereich konstruiert. Beim ZLT-II liest das Kind von Testkarten Einzellaute und Lautverbindungen, Wörter oder kurze Textabschnitte ab. Der Versuchsleiter notiert auf dem Testbogen die Art der Lesefehler und misst beim Lesen der Textabschnitte die dafür benötigte Zeit. Zusätzlich kann die phonologische Verarbeitung überprüft werden. Das Testmanual enthält zudem das Zusatzverfahren „Psycholinguistische Verlesungsanalyse“ und weitere Ausführungen zur Förderdiagnostik. Abhängig von der Klassenstufe dauert die Durchführung 15 bis 35 Minuten.

Für die einzelnen Subtests liegen die internen Konsistenzen zwischen $\alpha = .24$ und $\alpha = .91$, was zum Teil bedenklich niedrig ist. Für die einzelnen Klassenstufen werden jedoch zufriedenstellende Gesamtreliabilitäten von $\alpha = .83$ bis $\alpha = .93$ berichtet. Die Retestkoeffizienten liegen für die Lesegeschwindigkeiten zwischen $r = .93$ und $r = .99$, für die Lesefehler zwischen $r = .41$ und $r = .93$ und sind somit für die Lesegeschwindigkeiten sehr gut, für die Lesefehler zum Teil sehr niedrig. Die Norm-

stichprobe umfasst die Daten von 1 145 Kindern der ersten bis achten Klasse aus nur vier Regionen Deutschlands. Für jede Klassenstufe werden T-Werte und Prozentränge angegeben. Die Größe der Substichproben pro Klassenstufe liegt zwischen $n = 62$ und $n = 195$. Die Validität des ZLT-II wurde für Schüler mit einer anderen Muttersprache als Deutsch geprüft, wobei sich mit Ausnahme des Subtests „Schnelles Benennen“ keine bedeutsamen Unterschiede zu Kindern mit deutscher Muttersprache zeigten. Das spricht für eine Einsetzbarkeit des Tests bei Kindern mit einer anderen Muttersprache als Deutsch. Zwischen Schülern mit und ohne LRS kann der Test differenzieren. Die Korrelationen der Leseleistung mit den Schulnoten im Fach Deutsch bzw. im Lesen werden von den Autoren als hoch bezeichnet, fallen jedoch in der Sekundarstufe deutlich niedriger aus als in der Grundschule.

Insgesamt fallen die Ergebnisse für den ZLT-II für den Grundschulbereich sehr gut aus, für die Sekundarstufe sind die Ergebnisse allerdings weniger gut, was vor allem auf die mangelnde Differenzierungsfähigkeit einzelner Subtests in höheren Klassenstufen zurückzuführen sein dürfte. Durch den hohen Aufwand von Individualtestungen erscheint der ZLT II für den Schulalltag eher weniger geeignet, zudem sind die Normstichproben auf Klassenstufenebene recht klein.

Salzburger Lese-Screening 5-8 (SLS 5-8). Das SLS 5-8 (Auer et al., 2005) wurde zur Beurteilung der basalen Lesefertigkeit von Schülern der fünften bis achten Klassenstufe konstruiert. Bei diesem Test besteht die Aufgabe der Schüler darin, eine Liste von 70 inhaltlich einfachen und auf das Wissen der Schüler abgestimmten Sätzen möglichst schnell zu lesen und den Wahrheitsgehalt jedes Satzes mit einer entsprechenden Markierung am Ende der Zeile zu beurteilen. Die Anzahl der innerhalb von drei Minuten korrekt beurteilten Sätze bildet das Testergebnis. Das SLS 5-8 liegt in zwei Versionen mit vergleichbaren Anforderungen vor. Dadurch besteht die Möglichkeit, den Test sogar in kurzen Zeitabständen zu wiederholen. Für beide Satzversionen existieren zudem zwei Varianten mit veränderter Satzabfolge. Mit der Verwendung der beiden Abfolgevarianten kann im Klassenverband das Abschreiben vom Banknachbarn verhindert werden.

Der Test ist mit einer durchschnittlichen Paralleltestkorrelation von $r = .89$ reliabel. Die Korrelation mit der Leistung beim lauten Lesen von Texten liegt bei $r = .78$, was für eine valide Messung spricht. Der Test differenziert von der fünften bis achten Klassenstufe im gesamten Leistungsbereich. Für die vier Klassenstufen liegen separate Normen mit Stichprobenumfängen von $n = 714$ bis $n = 850$ vor. Die Gesamtstichprobe stammt hauptsächlich aus dem österreichischen Bundesland Salzburg und dem deutschen Bundesland Bayern. Eine Generalisierbarkeit auf den gesamten deutschen Sprachraum muss demnach auch hier nicht zwangsläufig gegeben sein.

Das SLS 5-8 ist als Gruppentest einsetzbar, die reine Bearbeitungszeit beträgt drei Minuten, die Durchführungzeit ca. zehn Minuten. Die Auswertung erfolgt recht schnell mithilfe von Schablonen. Somit ist der Test besonders ökonomisch und greift auf ei-

ne Methode zurück, die sich zur Erfassung des Leseverständnisses bewährt hat (vgl. W. Lenhard, 2013, S. 90). Beim SLS 5-8 handelt es sich allerdings um ein Screening, das nur basale Lesekompetenzen, nicht jedoch höhere Verständnisleistungen, überprüft.

Lesegeschwindigkeits- und -verständnistest für die Klassen 6-12 (LGVT 6-12). Der LGVT 6-12 (Schneider, Schlagmüller & Ennemoser, 2007) soll der Ermittlung des Leseverständnisses und der Lesegeschwindigkeit in den Klassenstufen sechs bis zwölf dienen. Er ist als Gruppen- oder Individualtest für Förderdiagnosen einsetzbar. Bei der Testung sollen die Schüler einen 1 727 Wörter umfassenden Fließtext möglichst schnell und möglichst genau lesen und versuchen, innerhalb von vier Minuten möglichst weit zu kommen. Die Zeit ist so knapp, dass es kaum möglich ist, den Text komplett zu lesen. An 23 über den Text verteilten Stellen sollen die Schüler aus jeweils drei Optionen das in den Textzusammenhang passende Wort auswählen.

Die Retestkoeffizienten (Intervall: 6 Wochen) liegen für die Leseverständnisskala (LGVT-LV) bei $r = .87$ ($N = 103$) und die Lesegeschwindigkeitsskala (LGVT-LG) bei $r = .84$ ($N = 103$). Beides kann also zuverlässig erfasst werden. Die LGVT-LV-Ergebnisse korrelierten mit den PISA 2000-Lesetestergebnissen zu $r = .59$ ($N = 711$) und mit den Ergebnissen des im nächsten Abschnitt erläuterten Lesestrategie-Wissenstests WLST 7-12 zu $r = .46$ ($N = 809$), was beides bedeutsam ist und für eine valide Messung spricht (Schneider, 2009; Schlagmüller & Schneider, 2007). Die LGVT-LG-Ergebnisse korrelierten mit den oben erwähnten Testergebnissen erwartungskonform deutlich niedriger, aber auch signifikant (mit den PISA 2000-Lesetestergebnissen zu $r = .35$ bei $N = 711$ und mit den WLST-Ergebnissen zu $r = .25$ bei $N = 809$). Prozentrangnormen für die Anzahl korrekter Unterstreichungen und die Anzahl gelesener Wörter liegen für die Klassenstufen sechs bis neun aller Schularten (außer Sonderschule) vor und für die zehnte Realschul- sowie die zehnte und elfte Gymnasialklasse. Die Normstichprobe besteht aus 2 390 Schülern aus elf deutschen Bundesländern. Mit einer Durchführungsdauer von ca. zehn Minuten und einer schnellen Auswertbarkeit kann der LGVT 6-12 als sehr ökonomisch angesehen werden. Die Nachteile der Lückentext-Methode wurden bereits mehrfach erläutert (vgl. Kap. 5.3.3 und Abschnitt zum VSL in diesem Kapitel).

Würzburger Lesestrategie-Wissenstest für die Klassen 7-12 (WLST 7-12). Der WLST 7-12 (Schlagmüller & Schneider, 2007) dient der Erfassung des Lesestrategiewissens von Schülern der Klassenstufen sieben bis zwölf. Er kann als Gruppen- und Individualtest zur Erstellung von Förderdiagosen eingesetzt werden. Beim WLST 7-12 werden sechs verschiedene Lernszenarien vorgelegt und die Schüler sollen für jedes Szenario die Qualität und Nützlichkeit von fünf verschiedenen Strategien zur Erreichung eines Lernziels bewerten.

Hohe Werte des Split-Half-Koeffizienten ($r = .88$), der internen Konsistenz ($r = .88$) und des Retestkoeffizienten (Intervall: 6 Wochen, $r = .81$) sprechen für reliable Mes-

sungen. Die mittelhohen Korrelationen der WLST-Ergebnisse mit denjenigen des PISA 2003-Lesetests ($r = .40$; $N = 3\,386$), eines Schnellleseverständnistests ($r = .46$; $N = 809$) und des Intelligenztests KFT ($r = .41$; $N = 3\,386$) sprechen für recht valide Messungen. Es liegen Prozentrangnormen für die aus Paarvergleichen erzielten Rohwertpunkte vor. Die Normstichprobe umfasste $N = 4\,490$ Schüler aus allen deutschen Bundesländern, was sehr positiv zu bewerten ist. Mit einer Durchführungsdauer von 20 bis 35 Minuten und der Einsetzbarkeit als Gruppentest ist die Durchführung des WLST 7-12 ökonomisch.

Leseverständnistest für Erwachsene (LEVE). LEVE (Proyer, Wagner-Menghin & Grafinger, 2010) ist ein im Rahmen des Wiener Testsystems als Online-Test verfügbarer Test zur Erfassung des Leseverständnisses. Er wurde auf der Grundlage der IRT entwickelt und ist Rasch-homogen. Der Test ist für Jugendliche (ab 12 Jahren) und für Erwachsene geeignet. Als Hauptanwendungsgebiete nennen die Autoren Personalpsychologie, Verkehrspsychologie und Pädagogische Psychologie. Beim LEVE wird am Bildschirm eine Geschichte dargeboten, die sich über mehrere Bildschirmseiten erstreckt, die der Proband selbständig umblättern muss. Zurückblättern ist nicht möglich. Am Ende müssen MC-Fragen zur gelesenen Geschichte beantwortet werden. Es stehen zwei Parallelversionen mit je einer Geschichte zur Verfügung. Es werden Kennwerte für die Anzahl vollständig gelöster Aufgaben, die Anzahl richtig gewählter Antwortmöglichkeiten, die Anzahl korrekterweise nicht gewählter Antwortmöglichkeiten und die gesamte Lesezeit ermittelt.

Die Autoren geben eine zufriedenstellende interne Konsistenz von $\alpha = .81$ für die eine Testform und gerade noch zufriedenstellende interne Konsistenz von $\alpha = .72$ für die andere Testform an. Aufgrund des Aufgabenprinzips nehmen die Autoren inhaltliche Validität an. Ergebnisse aus Extremgruppenvalidierungen zeigen, dass sich die Testleistungen von Personen mit Berufen, die ein hohes Maß an Leseverständnis erfordern, gegenüber jenen von Personen, die einen Beruf haben, für dessen Ausübung diese Kompetenz nicht unbedingt von Bedeutung ist, signifikant unterscheiden. Normdaten liegen von einer Stichprobe von $N = 392$ Personen für beide Texte vor, wobei zusätzlich drei altersspezifische Substichproben verfügbar sind. Die Daten wurden 2008 in Wien erhoben.

Als problematisch kann hier angesehen werden, dass die Normstichprobe mit $N = 392$ bei einer sehr großen Altersspanne recht klein ist. Darüber hinaus wurden die Normdaten ausschließlich in Wien erhoben, wodurch eine Übertragbarkeit auf andere deutschsprachige Gebiete nicht zwangsläufig gegeben sein muss. Bezüglich der Validität ist anzumerken, dass die Testpersonen nur eine Geschichte lesen. Somit beschränkt sich der Test auf lediglich ein Genre. Dies ist insofern problematisch, als beispielsweise in den meisten Berufen Sachtexte eine größere Rolle spielen dürfen, womit die Validität für die oben genannten Hauptanwendungsgebiete fraglich erscheint.

Kritische Bewertung der vorgestellten Tests. Insgesamt zeigt sich, dass bereits zahlreiche Tests zur Erfassung verschiedener Aspekte von Leseverständnis vorliegen. Die meisten sind jedoch nur für den Einsatz bis zur sechsten Klasse ausgelegt. Ab der siebten Klasse stehen deutlich weniger Tests zur Verfügung. Diese sind wie z. B. der ZLT-II als Individualtests wenig ökonomisch und für den Einsatz im Schulalltag wenig geeignet und/oder es handelt sich um Screenings, die nur basale Lesekompetenz erfassen wie z. B. das SLS 5-8, was deren Nützlichkeit für höhere Gymnasialklassen einschränkt. Sowohl der hohe Aufwand von Individualtests als auch die mangelnde Differenzierungsfähigkeit im oberen Leistungsbereich schränken darüber hinaus die Einsetzbarkeit für viele Forschungsfragestellungen ein.

Fast alle Tests sind jedoch als reliabel und valide zu bezeichnen. Lediglich der ZLT-II weist hier zum Teil bedenklich niedrige Werte auf. Kritisch ist zudem auch anzumerken, dass nur bei wenigen Tests Parallelformen existieren, was für wiederholte Messungen wie z. B. zur Erfassung des Lernfortschritts oder zur Evaluation von Fördermaßnahmen wünschenswert wäre. Die meisten Tests wurden auf der KTT basierend konstruiert und bieten Vergleichsnormen an, anhand derer erzielte Testergebnisse leichter beurteilt werden können. Die Normstichproben sind allerdings häufig bedenklich klein und stammen teilweise aus nur wenigen Regionen des großen, heterogenen deutschen Sprachraums, wodurch die Repräsentativität zweifelhaft erscheint. Niveaustufen werden beispielsweise beim FLVT 5-6 zur Verfügung gestellt.

5.3.5 PISA-Lesetests

Während die in den vergangenen Abschnitten beschriebenen Tests, die im Handel erhältlich oder online verfügbar sind, vor allem einzelne Aspekte von Lesekompetenz erfassen (insbesondere Dekodierfähigkeit, Leseverständnis oder Strategiewissen), wurden für die in den letzten Jahrzehnten im Bildungswesen vermehrt durchgeführten nationalen und internationalen Schulleistungsstudien Lesetests konzipiert, die vorwiegend auf dem Reading Literacy-Konzept basieren. Das Ziel dabei ist es, herauszufinden, ob die Lesekompetenz der Schüler ausreicht, um am informations- und kommunikationsorientierten gesellschaftlichen Leben teilhaben zu können, und ob bestimmte Bildungsziele erreicht werden. Als Beispiel wird im Folgenden die Konzeption der Lesetests von PISA betrachtet, da in der vorliegenden Arbeit häufig PISA-Befunde aufgegriffen werden.

PISA versucht, realitätsnahe Leseanforderungen in vielfältigen Kontexten abzubilden (Artelt, Drechsel, Bos & Stubbe, 2008). Die Anforderungen sind daher umfassender angelegt als in schulischen Curricula. Kognitionspsychologische Theorien von van Dijk und Kintsch (1983) sowie W. Kintsch (1994, 1998) und die sogenannte „Document Literacy“ nach Mosenthal und Kirsch (1991) bzw. I. S. Kirsch, Jungeblut und Mosenthal (1998) bilden die Grundlage für die Lesetests bei PISA (Schnotz & Dutke, 2004; Hurrelmann, 2006). Die Document Literacy bezieht nicht nur Texte, sondern

auch Diagramme, Tabellen und andere Formen schriftlicher Dokumente ein und besagt, dass sich die Verstehensanforderungen eines Dokuments in erster Linie durch dessen strukturelle Komplexität, die jeweiligen Aufgabenanforderungen und die sich daraus ergebenden notwendigen Lösungsoperationen ergeben. Hinsichtlich der Verstehensleistung wird vor allem dahingehend differenziert, ob alle für die Beantwortung nötigen Informationen textimmanent sind oder ob über den expliziten Textinhalt hinaus eine auf eigenem Vorwissen basierende Interpretation zur Beantwortung der Verständnisfragen erforderlich ist (Schnotz & Dutke, 2004). Um die so konzeptualisierte Kompetenz zu überprüfen, wurden unterschiedliche Textformen, z. B. ein deskriptiver Text, eine Erzählung oder ein Diagramm, in verschiedenen Kontexten dargeboten, z. B. in Form amtlicher Dokumente oder in Form von Lehrbuchtexten. Bei PISA 2009 wurden 29 Aufgabeneinheiten vorgegeben (Naumann, Artelt, Schneider & Stanat, 2010). Das Antwortformat war etwa zur Hälfte gebunden in Form von MC-Aufgaben und zur Hälfte offen (Jude & Klieme, 2010).

Die Gesamtskala der Lesekompetenz wurde jeweils so gebildet, dass die Ergebnisse aller PISA-Erhebungen unmittelbar vergleichbar sind (Naumann et al., 2010). Als Referenz dient die Gesamtskala von PISA 2000 mit einem Mittelwert vom 500 und einer Standardabweichung von 100 für die damaligen Teilnehmerstaaten (Naumann et al., 2010). Die Reliabilität der Lesekompetenzskala bei PISA 2009 liegt sowohl international als auch national bei $r_{tt} = .92$ (Naumann et al., 2010). Die kontinuierliche Fähigkeitsskala wurde in mehrere Niveaustufen unterteilt, und für die einzelnen Niveaustufen wurden die charakteristischen Aufgabenanforderungen inhaltlich beschrieben (Artelt et al., 2008). PISA 2000 bis PISA 2006 unterschieden fünf Niveaustufen. Schüler am unteren Ende einer Niveaustufe lösten durchschnittlich 62 % der leichten und 42 % der schweren Aufgaben dieser Stufe. Wer am oberen Ende der Stufe liegt, meisterte 78 % der leichten und 68 % der schweren Aufgaben. Für PISA 2009 wurden die Niveaustufen am oberen und unteren Ende erweitert, um in den Extrembereichen noch weiter differenzieren zu können (Naumann et al., 2010). Tabelle 1 beschreibt die Niveaustufen der Gesamtskala bei PISA 2009.

In den Jahren 2000 und 2009 wurde bei PISA die Lesekompetenz als Hauptdomäne erfasst. Die Auswertung erfolgte nicht nur auf einer Gesamtskala, sondern auch auf Subskalen. Die Subskalen beziehen sich auf unterschiedliche Aufgabenanforderungen („Aspekte des Lesens“) und unterschiedliche „Textformate“. Jede Aufgabe ist dabei einem Aspekt und einem Format zugeordnet. Bei den *Aspekten des Lesens* werden drei Verstehensanforderungen unterschieden: (1) Informationen in einem Text auffinden und extrahieren, (2) textbasiert Informationen kombinieren und interpretieren und (3) über Inhalt und Form eines Textes reflektieren (Artelt et al., 2008; Naumann et al., 2010). Bezüglich der *Textformate* wurde zwischen kontinuierlichen und diskontinuierlichen Texten unterschieden. Diskontinuierliche Texte enthalten bildhafte Darstellungen, z. B. Diagramme, Listen, Karten oder Tabellen, aus denen Informationen abgelesen und mit einem Text verknüpft werden müssen (Artelt et al., 2008; Naumann

et al., 2010). Kontinuierliche bzw. fortlaufende Texte bestehen aus in Absätzen organisierten Sätzen, wie dies bei Sachtexten und Prosa üblich ist. Kontinuierliche Texte sind z. B. als Erzählungen, Beschreibungen, Kommentare oder Argumentationen geschrieben. Zusätzlich wurden bei der Aufgabenkonstruktion verschiedene Leseanlässe berücksichtigt. Die verschiedenen *Situationen und Kontexte* bildeten zwar keine eigenen Kompetenzskalen, aber sollten für eine ausreichend breite Ausrichtung der Aufgaben sorgen (Naumann et al., 2010). Dabei wurde zwischen privat (z. B. Romane), öffentlich (z. B. Anschreiben), berufsbezogen (z. B. Handbücher) und bildungsbezogen (z. B. Lehrbücher) unterschieden (Artelt et al., 2008). Weiter wurden verschiedene

Tabelle 1. Niveaustufen des PISA 2009-Lesetests (nach Naumann et al., 2010, S. 28).

Stufe	Anforderungen
VI	Detailgenaues und präzises Schlussfolgern, Vergleichen und Gegenüberstellen unter Verwendung vollen und detaillierten Verständnisses eines oder mehrerer Texte, ggf. gedankliches Verknüpfen von Informationen aus mehreren Texten; Auseinandersetzung mit ungewohnten Ideen, kompetenter Umgang mit konkurrierenden Informationen und abstrakten Informationskategorien, präziser Umgang mit z. B. unauffälligen Textdetails
V	Finden und Ordnen mehrerer tief eingebetteter Informationen und Beurteilung der Relevanz dieser Information; auf Fachwissen basierende kritische Beurteilung oder Hypothesengenerierung; volles, detailliertes Verständnis von Texten ungewohnten Inhalts oder Form; Umgang mit erwartungswidrigen Konzepten
IV	Folgen linguistischer oder thematischer Verknüpfungen in einem Text über mehrere Abschnitte, z. T. ohne Verfügbarkeit eindeutiger Kennzeichen im Text, zum Finden, Interpretieren und Bewerten eingebetteter Informationen oder zur Erschließung psychologischer oder philosophischer Bedeutungen; genaues Verständnis langer oder komplexer Texte ungewohnten Inhalts oder ungewohnter Form
III	Nutzen vorhandenen Wissens über Textorganisation und -aufbau; Erkennen impliziter oder expliziter logischer Relationen über mehrere Sätze oder Textabschnitte zur Lokalisation, Interpretation oder Bewertung von Texten; Begreifen eines Zusammenhangs oder Analyse eines Wortes oder Satzes, wobei benötigte Informationen nicht immer leicht sichtbar sind und Textpassagen häufig erwartungswidrig verlaufen
II	Folgen logischer und linguistischer Verknüpfungen innerhalb eines Textabschnitts zur Lokalisation oder Interpretation von Informationen; Verknüpfen von im Text oder über Textabschnitte verteilten Informationen zur Erschließung der Autorenintention; Vergleiche und Gegenüberstellungen auf Basis eines einzigen Textbestandteils; Vergleiche ausgehend von eigenen Erfahrungen oder Standpunkten; Erkennen von Zusammenhängen zwischen im Text enthaltenen und nicht enthaltenen Informationen
Ia	Lokalisation explizit ausgedrückter Informationen in einem Text zu einem vertrauten Thema; Erkennen des Hauptthemas oder der Autorenintention oder einfacher Zusammenhänge zwischen textimmanenten Informationen und Alltagswissen; die Informationen sind leicht zu finden; wenige oder keine Informationen konkurrieren; Hinweise auf die entscheidenden Elemente der Aufgabe und des Textes werden explizit gegeben.
Ib	Lokalisation einer einzigen, explizit ausgedrückten und leicht sichtbaren Information in einem kurzen, syntaktisch einfachen Text vertrauter Form und aus einem gewohnten Kontext; Text enthält i. d. R. Hilfestellungen (Wiederholungen, Bilder, bekannte Symbole); kaum konkurrierende Informationen; Herstellen einfacher Zusammenhänge zwischen benachbarten Informationen

Funktionen und Zwecke berücksichtigt, nämlich Beschreibung, Erzählung, Darlegung, Argumentation, Anleitung und Transaktion (Naumann et al., 2010).

Trotz der sorgfältigen Konstruktion der Skalen erfuhr die zugrundegelegte Konzeption von Lesekompetenz einige Kritik. Eher didaktisch orientierte Autoren bemängelten, dass die Konzeption zu stark auf die kognitiven Aspekte fokussiere und zu wenig ganzheitlich sei, wichtige Aspekte (z. B. emotional-motivational oder sozial) blieben unberücksichtigt (vgl. Hurrelmann, 2002; Groeben, 2004; Rosebrock & Nix, 2011). Damit reiche diese Konzeption zwar für das Messen der Leseverstehensleistung aus, aber nicht für lesedidaktische Zwecke (Hurrelmann, 2006; Rosebrock & Nix, 2011). Die PISA-Autoren weisen jedoch explizit darauf hin, dass es nicht Ziel von PISA war, den elementaren Bereich sprachlich-literarischer Allgemeinbildung zu erfassen (Baumert et al., 2001, S. 21). Vielmehr wollte PISA anhand reliabler und valider Messungen individuelle Leistungen in Relation zu Variablen der verschiedenen Bildungssysteme setzen. Dabei ist es nachrangig, die der Lesekompetenz zugrunde liegenden inneren Strukturen und Prozesse zu erfassen, stattdessen steht die psychometrische Analyse von Kompetenzunterschieden im Vordergrund. Eine Erweiterung der Konzeption von Lesekompetenz für PISA 2009 trägt der Kritik dennoch ein Stück weit Rechnung, indem die aktive Lesepraxis, die ein gewisses Maß an Motivation voraussetzt, mitberücksichtigt wird (Artelt & Stanat, 2010). Reading Literacy innerhalb von PISA wird nicht mehr nur als Verstehen, Nutzen und Reflektieren schriftlicher Texte verstanden, sondern auch als das sich auf das Lesen einlassen, um eigene Ziele zu erreichen, als das Erweitern von Wissen und Kompetenzen und das Teilnehmen an der Gesellschaft (Artelt & Stanat, 2010). Das „Reading Engagement“ ist dabei recht weit gefasst und schließt neben der Lesemotivation auch soziale Aspekte sowie Lese- und Textverstehensstrategien ein (Artelt & Stanat, 2010).

Weitere Kritik an der Konzeption der PISA-Lesetests bezieht sich auf die Zugrundelegung eines ausdifferenzierten Lesekompetenzmodells, das nicht dem aktuellen Forschungsstand entspricht (D. H. Rost & Buch, 2010). Es wurden fünf Skalen voneinander abgegrenzt. Diese Ausdifferenzierung wurde jedoch empirisch nicht bestätigt. Stattdessen ergaben sich die drei Leseverständnisfacetten „Informationen ermitteln“, „textbezogene Interpretation“ sowie „Reflektieren und Bewerten“. Da diese Facetten aber sehr hoch korrelieren, sind sie eher als Subskalen *einer* Kompetenz zu betrachten denn als distinkte Skalen (vgl. D. H. Rost & Buch, 2010; Schaffner, Schiefele & Schneider, 2004; Artelt et al., 2001). Laut Schnotz und Dutke (2004, S. 97) liegt ein wesentlicher Erfolg von PISA jedoch gerade darin, Testaufgaben geschaffen zu haben, die sich empirisch abgesichert auf einer eindimensionalen Schwierigkeitsskala anordnen lassen und damit einen zuverlässigen Vergleichsmaßstab für Leseverstehensleistungen darstellen. Mithilfe der IRT gelang eine reliable und inhaltlich valide Messung, die weitgehend den Leseanforderungen vieler Menschen im heutigen Alltag entspricht. Die Aufgaben stellen multiple Anforderungen an den Aufbau mentaler Repräsentationen und sind damit exemplarisch für einen Begriff von Lesen und Verste-

hen, der relativ gut mit den kognitionspsychologischen Forschungserkenntnissen zum Text- und Bildverstehen übereinstimmt (Artelt et al., 2001; Schnotz & Dutke, 2004).

5.4 Fazit

Im vorliegenden Kapitel wurde nach einer Darstellung möglicher klinischer Diagnosen im Zusammenhang mit dem Lesen deutlich, dass Lehrkräfte zum Teil große Schwierigkeiten haben, Defizite in der Lesekompetenz bei ihren Schülern zu erkennen. Es zeigte sich, dass die diagnostische Kompetenz von Lehrkräften sowohl intra- als auch interindividuell stark variiert, dass zur Beurteilung der Schülerleistung in der Regel nur die einzelne Klasse als Referenzrahmen herangezogen wird und dass Schulnoten nach sehr unterschiedlichen Maßstäben vergebenen werden. Lehrkräfte weiterführender Schulen können die Lesekompetenz ihrer Schüler weniger akkurat einschätzen als Grundschullehrkräfte, und die Leseleistung wird tendenziell eher überschätzt. Auch bezüglich der Identifikation besonders schwacher Leser sowie der den schwachen Leseleistungen zugrunde liegenden Defizite scheinen Probleme zu bestehen. Diese Erkenntnisse sprechen dafür, dass standardisierte und normierte Lesetests Lehrkräfte und insbesondere Lehrkräfte weiterführender Schulen bei der Beurteilung der Lesekompetenz ihrer Schüler unterstützen könnten, indem ein erweiterter Referenzrahmen zur Verfügung gestellt und ein differenziertes Bild der Leseleistung erstellt wird, das Rückschlüsse auf die mangelndem Leseverständnis zugrunde liegenden Defizite zulässt.

Solche Tests sollten stets auf einer Theorie basieren, welche das Ziehen von Schlüssen aus dem Verhalten einer Person in einer Testsituation auf die Ausprägung des interessierenden Merkmals bei dieser Person ermöglicht. Hierfür existieren zwei Theorien, die im vorliegenden Kapitel erläutert wurden: Die KTT und die IRT. Die KTT legt den Schwerpunkt auf Reliabilitätsaspekte und ermöglicht die Bestimmung der Reliabilität eines Tests. Darauf basierend lässt die KTT Rückschlüsse auf die wahre Ausprägung des interessierenden Merkmals zu. Die IRT dagegen behandelt vor allem die Skalierbarkeit und die Konstruktvalidität. Beide Theorien lassen sich jedoch kombinieren und können sich gegenseitig ergänzen. Unabhängig von der zugrunde gelegten Theorie gibt es allgemein anerkannte Gütekriterien, denen psychologische Tests genügen sollten. Über die Hauptgütekriterien der Objektivität, Reliabilität, Validität und Skalierbarkeit hinaus wurden weitere, sogenannte Nebengütekriterien, wie z. B. Normierung und Ökonomie, betrachtet. Es ist ein großer Vorteil von standardisierten Tests, dass sich ihre Qualität anhand der Angaben zu den Gütekriterien beurteilen lässt. Ein weiterer Vorteil standardisierter Tests liegt darin, dass sie es erlauben, ohne Vorkenntnisse über ein Kind eine Einschätzung der Leistung des Kindes zu erhalten. Dies kann z. B. für Lehrkräfte bei der Übernahme einer neuen Klasse eine große Hilfe sein (W. Lenhard, 2013, S. 70). Zudem ist in der Regel eine Bezugsnorm gegeben, die einen

Vergleich der Schülerleistungen mit einer Referenzgruppe ermöglicht. Standardisierte Tests sind darüber hinaus bezüglich Durchführung, Auswertung und Interpretation standardisiert und liefern objektive Ergebnisse, was einer verzerrten Beurteilung vorbeugt. Die Tests können daher bei sozialrechtlichen oder juristischen Entscheidungen sowie Entscheidungen über die Schullaufbahn als objektive Ergänzung zu subjektiven Beurteilungen z. B. durch Lehrkräfte und Eltern dienen und führen zu einer faireren Entscheidung bezüglich Selektion und Fördermaßnahmen. Nicht zuletzt basieren standardisierte Tests meist auf dem aktuellen Forschungsstand und einer entsprechenden Theorie, die in einem dazugehörigen Testmanual explizit erläutert wird.

Es gibt jedoch auch einige Einschränkungen standardisierter Tests. So prüfen sie in der Regel eher grundlegende Fähigkeiten und nicht das Erreichen eines bestimmten Lernziels; sie haben also keinen konkreten Unterrichtsbezug (W. Lenhard, 2013, S. 70). Es handelt sich bei standardisierten Tests zudem größtenteils um normierte Tests, die zwar einen sozialen Vergleich relativ zu einer Referenzgruppe ermöglichen, aber keine Beurteilung im Hinblick auf absolute Kriterien. IRT-skalierte Tests haben hier zumindest prinzipiell die Möglichkeit, anhand von Niveaustufen auch eine inhaltliche Beurteilung zuzulassen, wobei diese Möglichkeit auch umstritten ist, was im entsprechenden Abschnitt erläutert wurde. Ein weiteres Problem normierter Tests ist die Frage nach der Qualität der Normen. So ist z. B. fraglich, ob es möglich ist, eine tatsächlich repräsentative Normstichprobe zu ziehen (für eine Übersicht bzgl. der Schwierigkeit, eine repräsentative Stichprobe zu ziehen s. Wyschkon, 2011). Auch die schnelle Veraltung von Normen ist kritisch zu sehen, da eine Normierung nach acht bis maximal zehn Jahren ihre Gültigkeit verliert.

Generell können Lesetests auf unterschiedliche Aspekte oder Teilprozesse des Lesens fokussieren, wobei es unterschiedliche Möglichkeiten gibt, diese zu operationalisieren. Die unterschiedlichen Varianten wirken sich dabei z. B. auf die Ökonomie des Tests aus (Zeitaufwand, reiner Individualtest vs. Einsatz als Gruppentest möglich). Ein Überblick über aktuelle, gängige Lesetests für die Sekundarstufe zeigte weiter, dass inzwischen einige Tests zur Verfügung stehen, die größtenteils den Gütekriterien genügen. Als problematisch erwiesen sich jedoch meist die Normstichproben hinsichtlich Umfang und Repräsentativität sowie die Ökonomie bei jenen Tests, die ausschließlich als Individualtests eingesetzt werden können. Zudem beschränken sich die als Gruppentests einsetzbaren Tests ab der siebten Klassenstufe bislang eher auf basale Leseprozesse und erlauben keine Differenzierung im oberen Leistungsbereich. Ein Test, der in den mittleren und höheren Klassenstufen eingesetzt werden kann und der im gesamten Leistungsspektrum differenziert, als Gruppentest einsetzbar ist und über Normdaten einer großen, annähernd repräsentativen Stichprobe verfügt, fehlt bisher. Idealerweise würde ein solcher Test auch über eine Parallelversion verfügen und wäre bezüglich der Anwendung und Auswertung ökonomisch. Wünschenswert wäre weiter eine Konstruktion, die sowohl auf der KTT als auch auf der IRT basiert, um die Vorteile beider Testtheorien nutzen zu können (vgl. Kap. 5.3.1).

Genauer beschrieben wurde im vorliegenden Kapitel außerdem die Konzeption der PISA-Lesetests, denn auf diese auf dem angloamerikanischen Literacy-Konzept beruhenden Tests gehen zahlreiche der im folgenden Kapitel betrachteten Befunde zur Lesekompetenz deutscher Sekundarschüler zurück.

Kapitel 6

Aktuelle Befunde zur Lesekompetenz deutscher Sekundarschüler

Das folgende Kapitel beschäftigt sich mit aktuellen Befunden zur Lesekompetenz deutscher Sekundarschüler. Dabei werden zunächst Befunde zur kriteriumsorientierten Leseleistung bzw. zu den von den Schülern erreichten Niveaustufen betrachtet, bevor auf Unterschiede zwischen verschiedenen nationalen Subgruppen eingegangen wird, wobei die Bundesländer, die Schularten und die Geschlechter ebenso verglichen werden wie Schüler mit und ohne Migrationshintergrund.

6.1 Kriteriumsorientierte Beurteilung der Lesekompetenz

Am Ende der Grundschulzeit zeigten bei IGLU 2011 etwas weniger als 10 % der deutschen Schüler eine Leseleistung, die der höchsten Niveaustufe zuzuordnen ist, während mehr als 15 % nicht einmal eine Niveaustufe erreichten, die als ausreichend bezeichnet werden kann (Tarelli, Valtin, Bos, Bremerich-Vos & Schwippert, 2012). Letztere stellen damit langfristig eine Risikogruppe für gravierende Lernprobleme dar. Insgesamt fiel die Leseleistung der deutschen Viertklässler für literarische Texte besser aus als für Sachtexte. Darüber hinaus zeigte sich eine höhere Verständnisleistung bezüglich textimmanenter Aspekte, die genaues, gründliches und sorgfältiges Lesen erfordern, im Vergleich zu wissensbasierten Verständnisleistungen.

Bei DESI konnten 96 % der Schüler am Ende der neunten Klasse sinntragende Elemente im Text auffinden, 32 % waren in der Lage, zielgerichtet zu lesen und Informationslücken im Text zu schließen, 10 % konnten mit großer Sicherheit übergeordnete Textstrukturen erkennen und mit eigenem Vorwissen verknüpfen und 6 % konnten Situationsmodelle auswerten, also z. B. die Hauptfiguren und ihre Relationen, Ort, Zeit sowie ein zentrales Motiv zusammenfügen (vgl. DESI-Konsortium, 2006, S. 5f.). Für die überwiegende Mehrheit der Gymnasiasten schien es bei DESI keine große Herausforderung darzustellen, auch an schwierigen Textstellen durch genaues Lesen und Inferenzen Lücken in einem Text zu schließen. In den anderen untersuchten Schularten, insbesondere der Realschule, gelang es zwar vielen Schülern, angemessen fokussiert zu lesen – jedoch in den meisten Fällen nicht mit der erforderlichen Sicherheit, um der entsprechenden Niveaustufe zugeordnet zu werden. Denn die Zuordnung zu einer

Niveaustufe erfolgte nur dann, wenn ein Schüler die entsprechenden Anforderungen in zwei von drei Fällen bewältigen konnte. Die entscheidende Herausforderung schien bei Neuntklässlern darin zu bestehen, ein übergreifendes Verständnis herzustellen, bei dem im Text enthaltene Informationen, bereits vorhandenes Vorwissen und Inferenzen als Basis für eine eigenständige Interpretationsleistung zusammengeführt werden. Selbst in den Realschulen und integrierten Gesamtschulen wurden diese Fähigkeiten nur von wenigen Schülern mit der ausreichenden Zwei-Drittel-Sicherheit beherrscht (DESI-Konsortium, 2006, S. 6).

Bei PISA 2009 erreichten 18.5 % der deutschen 15-Jährigen nur eine der untersten Niveaustufen (Ia, Ib) oder gar keine, während 7.6 % die höchsten Niveaustufen (V, VI) erreichten (Naumann et al., 2010, Erläuterung der Niveaustufen s. Tab. 1 in Kap. 5.3.5). In Bezug auf die unterschiedlichen Textformate ergaben sich keine Leistungsunterschiede zwischen kontinuierlichen und diskontinuierlichen Texten. Es zeigte sich jedoch eine tendenzielle Schwäche in Bezug auf den Kompetenzaspekt Reflektieren/Bewerten im Vergleich zu den tendenziell eher stärkeren Aspekten Suchen/Extrahieren und Kombinieren/Interpretieren.

Insgesamt sprechen die Ergebnisse von DESI und PISA somit dafür, dass sich die bei IGLU bereits am Ende der Grundschulzeit gefundene relative Schwäche der deutschen Schüler für höhere Verständnisleistungen, die über den Textinhalt hinausgehen und das Hinzuziehen von Vorwissen erforderlich machen, auch in der Sekundarschulzeit fortzusetzen scheint.

6.2 Bundesländervergleich

Im Jahr 2009 wurden im Rahmen einer nationalen Bundesländervergleichsstudie des Instituts für Qualitätsentwicklung im Bildungswesen (IQB) alle 16 Bundesländer hinsichtlich der Lesekompetenz von Neuntklässlern verglichen (Köller, Knigge & Tesch, 2010). 36 605 Schüler aus 1 655 Klassen und 1 466 Schulen wurden getestet (Böhme et al., 2010). Die Schüler aus Bayern, Sachsen und Baden-Württemberg erzielten dabei signifikant über dem Bundesdurchschnitt liegende Werte, während die Leistungen der Schüler aus Brandenburg, Hamburg, Berlin und Bremen signifikant darunter lagen (s. Tab. 2). Dabei unterschieden sich die Bundesländer nicht nur im mittleren Leistungsniveau, sondern auch hinsichtlich der Streuung (Schipolowski & Böhme, 2010).

Die gesonderte Betrachtung der Leseleistung von Gymnasiasten zeigte, dass diese in Sachsen, Bayern, dem Saarland und Baden-Württemberg Ergebnisse erzielten, die signifikant über dem Bundesdurchschnitt der Gymnasiasten lagen (Schipolowski & Böhme, 2010). Die Gymnasiasten aus Hamburg, Mecklenburg-Vorpommern, Bremen, Hessen und Brandenburg lagen dagegen signifikant unter dem deutschen Mittelwert der Gymnasiasten. Dass der Unterschied zwischen acht- und neunjährigen Gymnasialmodellen für die Leistungsunterschiede verantwortlich war, konnte ausgeschlossen

werden. Bezüglich der übrigen Schularten konnten aufgrund der Heterogenität der Schulstrukturen in den 16 Bundesländern keine sinnvollen Vergleiche vorgenommen werden (Schipolowski & Böhme, 2010).

Tabelle 2. Mittelwert (M) und Standardabweichung (SD) im Lesetest der IQB-Studie zum Bundesländervergleich im Jahr 2009 (Schipolowski & Böhme, 2010).

Bundesland	***M***	***SD***	**Bewertung**
Bayern	509	89	überdurchschnittlich
Sachsen	508	97	
Baden-Württemberg	504	87	
Thüringen	497	87	durchschnittlich
Rheinland-Pfalz	497	92	
Sachsen-Anhalt	496	89	
Bundesdurchschnitt	*496*	*92*	
Mecklenburg-Vorpommern	493	88	
Saarland	492	93	
Hessen	492	90	
Nordrhein-Westfalen	490	98	
Niedersachsen	490	100	
Schleswig-Holstein	488	96	
Brandenburg	485	89	unterdurchschnittlich
Hamburg	484	99	
Berlin	480	105	
Bremen	469	104	

6.3 Unterschiede zwischen den Schularten

Da für die vorliegende Arbeit vor allem Befunde für deutsche Regelschulen relevant sind, wird beim folgenden Vergleich der Leseleistung zwischen den Schularten auf Befunde zur Leseleistung in Sonder- und Förderschulen nicht eingegangen.

In einer Studie von Retelsdorf und Möller (2008) bestanden schon zu Beginn der fünften Klasse deutliche Unterschiede in der Lesekompetenz zwischen allen drei untersuchten Schularten (Gymnasium, Realschule und Hauptschule). Eine weitere Studie von Retelsdorf und Möller (2011) verglich Gymnasiasten und Nicht-Gymnasiasten hinsichtlich Dekodiergeschwindigkeit und Leseverständnis. Dabei zeigten sich bereits zu Beginn der fünften Klasse deutliche Unterschiede zwischen Gymnasiasten und Nicht-Gymnasiasten für beide Komponenten der Lesekompetenz. Bezüglich der Dekodierleistung wurde zudem selbst bei Kontrolle verschiedener relevanter Hintergrundvariablen im weiteren Verlauf bis zur achten Klasse ein Schereneffekt dahingehend gefunden, dass der Leistungszuwachs bei den Gymnasiasten deutlich stärker ausfiel als bei

den Nicht-Gymnasiasten. Hinsichtlich des Leseverständnisses war jedoch kein schulartabhängiger Unterschied im Leistungszuwachs zu verzeichnen.

Bei PISA 2000 ließen sich im Bereich Lesekompetenz 45 % der Leistungsvarianz auf die Schulart zurückführen, bei DESI waren es im Bereich Deutsch 52 % (DESI-Konsortium, 2006, S. 54). Die Ergebnisse der deutschen Hauptschüler lagen bei PISA 2009 etwa zwei Standardabweichungen unter den Ergebnissen der Gymnasiasten. Die Realschüler und die Schüler der integrierten Gesamtschulen lagen dazwischen (Naumann et al., 2010). Der Mittelwert der Gymnasiasten lag auf Niveaustufe IV, der der Realschüler im unteren Drittel der Niveaustufe III, der der Gesamtschüler im oberen Bereich der Niveaustufe II und der der Hauptschüler auf Niveaustufe II. In der Hauptschule erreichte knapp die Hälfte der Schüler nicht Niveaustufe II, in der integrierten Gesamtschule waren dies ca. 18 %, in der Realschule ca. 10 % und am Gymnasium nur 0.5 % (Naumann et al., 2010). Demgegenüber fielen die Ergebnisse von 20 % der Gymnasiasten in den Bereich der obersten Niveaustufen (V und VI), während die anderen Schularten auf diesen Stufen nur mit sehr kleinen Anteilen vertreten waren, nämlich die Realschule mit 2.7 %, die integrierte Gesamtschule mit 1.4 % und die Hauptschule mit 0.2 %. Trotz der deutlichen Mittelwertsunterschiede bestand allerdings jeweils ein nicht zu vernachlässigender Überlappungsbereich zwischen den Schularten. Das obere Viertel der Hauptschüler erreichte bessere Ergebnisse als das untere Viertel der Realschüler. Der Überlappungsbereich zwischen Realschule und Gymnasium fiel ähnlich groß aus. Sogar zwischen Hauptschule und Gymnasium bestand ein kleiner Überlappungsbereich: Die 10 % besten Hauptschüler erzielten bessere Ergebnisse als die 10 % schlechtesten Gymnasiasten.

6.4 Geschlechterunterschiede

In der Hamburger Längsschnittstudie stieg der Geschlechterunterschied im Hinblick auf die Leseleistung zugunsten der Mädchen von der fünften zur neunten Klasse deutlich an (R. H. Lehmann, Peek, Gänsfuß & Husfeldt, 2002). In einer IEA-Lesestudie dagegen verlor sich der Vorsprung der Mädchen von der 3. zur 8. Klasse weitgehend. Zudem war der Leistungsnachteil der Jungen bei narrativen Texten ausgeprägter als bei Sach- und Gebrauchstexten (R. H. Lehmann, 1994). Bei S. Meyer (2009) zeigte sich bei Siebtklässlern kein Geschlechterunterschied in der Lesekompetenz. Betrachtet man die Normdaten verschiedener Lesetests für die Sekundarstufe, so zeigt sich, dass die Mädchen bei der Normierung des SLS 5-8 geringfügig besser abschnitten als die Jungen, und bei der Normierung des LGVT 6-12 ergaben sich kleine, jedoch nicht bedeutsame Geschlechterunterschiede zugunsten der Mädchen (Schneider, 2009).

Bei PISA schnitten Mädchen in der Lesekompetenz kontinuierlich signifikant besser ab als Jungen (Naumann et al., 2010). Bei PISA 2009 fiel der Geschlechterunterschied sogar noch etwas größer aus als zuvor. In den Bereich sehr schwacher Leistungen

(unter Niveaustufe Ia) fielen 3 % der Mädchen und 8 % der Jungen. In den Bereich schwacher Leistungen (unter Niveaustufe II) fielen ca. 13 % der Mädchen und 24 % der Jungen. Betrachtet man demgegenüber den oberen Leistungsbereich (Niveaustufe V und VI) findet man umgekehrte Verhältnisse. Mit etwa 11 % war der Anteil der Mädchen mehr als doppelt so hoch wie der Anteil von etwas mehr als 4 % der Jungen. In der mittleren Niveaustufe (III) war das Geschlechterverhältnis recht ausgeglichen. Signifikante Unterschiede bestanden bei PISA 2009 in allen erfassten Bereichen der Lesekompetenz. Auch in der DESI-Studie wurden in allen Kompetenzbereichen für die Domäne Deutsch statistisch bedeutsame Geschlechterunterschiede zugunsten der Mädchen gefunden. Für die Lesekompetenz fielen die Unterschiede jedoch im Gegensatz zu den anderen Sprachbereichen (außer Wortschatz) eher gering aus (DESI-Konsortium, 2006, S. 20).

Insgesamt fallen Leistungsunterschiede zwischen den Geschlechtern – sofern sie auftreten – also stets zugunsten der Mädchen aus. Häufig sind die Unterschiede signifikant, jedoch ist die Effektstärke in der Regel klein (z. B. bei PISA 2009: $d = 0.4$, Naumann et al., 2010) und die Effekte sind kaum praktisch bedeutsam (vgl. W. Lenhard, 2013, S. 60f.). Bei statistischer Kontrolle des Einflusses von Lesemotivation und Lesestrategiewissen unterschieden sich die Geschlechter auch bei PISA 2009 nicht mehr in der Leseleistung (Artelt, Naumann & Schneider, 2010). Auf Erklärungsansätze für das Zustandekommen der Geschlechterunterschiede in der Lesekompetenz wird in Kapitel 7 noch näher eingegangen.

6.5 Schüler mit vs. ohne Migrationshintergrund

Der Migrationshintergrund wird in verschiedenen Studien auf unterschiedliche Weise berücksichtigt. PISA definiert den Migrationsstatus beispielsweise über das Geburtsland der Eltern und der Schüler selbst, während DESI zwischen Schülern mit Deutsch als Erstsprache und Schülern mit einer anderen Erstsprache als Deutsch differenziert. Wieder andere Studien ziehen die Familiensprache als Kriterium heran (z. B. Gräsel et al., 2007). Allerdings zeigte sich, dass die über das Geburtsland definierten Gruppen weitgehend mit den Gruppen übereinstimmen, die nach dem Erstsprache-Kriterium eingeteilt wurden (Stanat, Rauch & Segeritz, 2010; DESI-Konsortium, 2006).

In einer Studie von Gräsel et al. (2007) fanden sich bereits in den ersten Sekundarschuljahren Unterschiede in der Lesegeschwindigkeit und im Leseverständnis zwischen Hauptschülern mit Deutsch als Familiensprache, Hauptschülern mit Deutsch und einer weiteren Sprache als Familiensprachen sowie Hauptschülern mit ausschließlich einer anderen Familiensprache als Deutsch. Erstgenannte zeigten im Hinblick auf beide Komponenten der Lesekompetenz die besten Ergebnisse, Zweitgenannte lagen im Mittelfeld und Letztgenannte zeigten die niedrigsten Lesekomeptenzwerte. Die Unterschiede zwischen den Gruppen waren gering, jedoch bestanden sie relativ stabil

über den Untersuchungszeitraum von zwei Schuljahren hinweg. Es fand sich weder eine Annäherung der Schüler mit Migrationshintergrund an die Schüler ohne Migrationshintergrund noch zeigte sich ein Schereneffekt.

Bei DESI schnitten ebenfalls Neuntklässler mit einer anderen Erstsprache als Deutsch in der Lesekompetenz schlechter ab als Neuntklässler mit Deutsch als Erstsprache (DESI-Konsortium, 2006). Schüler mit Deutsch und einer weiteren Erstsprache unterschieden sich nicht signifikant von Schülern mit nur Deutsch als Erstsprache. Allerdings war der Leistungsunterschied im Hinblick auf den Wortschatz noch deutlich größer, wobei hier auch mehrsprachige Schüler schlechtere Leistungen zeigten als Schüler mit nur Deutsch als Erstsprache.

Auch bei PISA fielen sowohl in Deutschland als auch in den meisten anderen Teilnehmerländern die Leistungen der 15-Jährigen mit Migrationshintergrund bezüglich der Lesekompetenz deutlich schlechter aus als die Leistungen der 15-Jährigen ohne Migrationshintergrund (z. B. Stanat et al., 2010). In Deutschland geborene Schüler, bei welchen für beide Elternteile Deutsch nicht die Muttersprache ist, schnitten bei PISA nicht viel besser ab als Schüler, die selbst im Ausland geboren wurden. War dagegen für ein Elternteil Deutsch die Muttersprache, schnitten Schüler mit Migrationshintergrund nur unwesentlich schlechter ab als Schüler ohne Migrationshintergrund. Bei PISA 2009 erzielten Schüler mit Migrationshintergrund deutlich bessere Ergebnisse als bei PISA 2000, was bei etwa gleichbleibender Leistung der Schüler ohne Migrationshintergrund zu einer Annäherung der beiden Gruppen führte. Diese Annäherung der Schüler mit Migrationshintergrund an jene ohne Migrationshintergrund lässt sich dabei nicht auf eine Verbesserung der sozialen Situation dieser Schüler zurückführen.

6.6 Fazit

Das vorliegende Kapitel stellte zunächst Befunde dar, die zeigten, dass bei deutschen Schülern der mittleren Klassenstufen eine relative Schwäche im Hinblick auf die höheren Verständnisebenen besteht. Dieses Ergebnis unterstreicht die zuvor bereits dargestellte Notwendigkeit, die bestehende Lücke in der Leseverständnisdiagnostik im Bereich der höheren Verständnisebenen für die mittleren und höheren Klassenstufen zu schließen. Standardisierte Tests, die die entsprechenden Aspekte des Leseverständnisses erheben, könnten den Lehrkräften helfen, die Schwächen zu identifizieren und durch entsprechende Fördermaßnahmen zu beheben.

Vergleiche verschiedener Subpopulationen ergaben, dass sich die Schüler der verschiedenen Bundesländer in ihrer Leseleistung deutlich unterscheiden. Zudem zeigte sich zwischen den Schularten ein bedeutsamer Unterschied, wobei erwartungskonform die Gymnasisten die besten Leistungen erzielen, gefolgt von den Real- und Gesamtschülern. Das Schlusslicht bilden die Hauptschüler. Zugleich gibt es jedoch zwischen allen Schularten auch Überlappungsbereiche. Weitere Leistungsunterschiede

finden sich häufig hinsichtlich des Geschlechts (zugunsten der Mädchen). Im Hinblick auf den Migrationshintergrund bzw. die Erst- oder Familiensprache fallen die Leseleistungen stets zugunsten der Schüler ohne Migrationshintergrund bzw. mit Deutsch als Erst- oder Familiensprache aus. Die Erkenntnis, dass zwischen diesen verschiedenen Subgruppen deutliche Leistungsunterschiede vorliegen, ist eine wichtige Information für die Normierung von Lesetests. Sowohl bei der Stichprobenziehung als auch bei der Beurteilung der Repräsentativität der Normstichproben ist die Verteilung der Schüler über die genannten Subgruppen zu beachten.

Nachdem in den vergangenen Abschnitten Differenzen zwischen verschiedenen Subgruppen von Schülern im Hinblick auf die Lesekompetenz betrachtet wurden, stellt sich nun die Frage, wie diese Leistungsdifferenzen zustande kommen. Einige Hinweise darauf wurden bereits bei der Betrachtung der Lesekompetenzentwicklung (s. Kap. 3) angesprochen, im folgenden Kapitel wird nun noch einmal vertieft auf die multiplen Einflussfaktoren eingegangen.

Kapitel 7

Einflussfaktoren auf das Leseverständnis

Interindividuelle Differenzen in der Lesekompetenz sind multifaktoriell bedingt und können alle Teilprozesse des Lesens betreffen und somit auf allen Ebenen des Leseverständnisses auftreten (vgl. Schneider, 2008). Die Forschung ist sich, wie bereits in Kapitel 4 deutlich wurde, weitgehend einig, dass die Hauptquellen interindividueller Differenzen bezüglich des Leseverständnisses in unterschiedlicher Ausprägung der basalen Lesekompetenz und der verbalen Intelligenz bzw. dem generellen rezeptiven Sprachverständnis liegen. Auch z. B. bei PISA erwies sich die verbale Intelligenz als bedeutendster Prädiktor für die Lesekompetenz (Artelt et al., 2001; Schaffner, Schiefele, Drechsel & Artelt, 2004).

Im Folgenden werden nun weitere Faktoren betrachtet, die die basale Lesekompetenz und das Leseverständnis beeinflussen. Dabei spielen Merkmale des Lesers, verschiedene Aspekte des sozialen Umfeldes sowie die Beschaffenheit des Textes eine Rolle. Die Bedeutsamkeit der einzelnen Faktoren ist noch nicht vollständig geklärt, denn es ist häufig schwierig, eindeutig festzustellen, ob beobachtete Phänomene Ursache oder Folge schwacher Leseleistungen sind oder ob sowohl die schwachen Leseleistungen als auch die anderen Phänomene durch eine dritte Variable verursacht werden (vgl. Berger, 2010, S. 40f.). Es zeichnet sich jedoch ab, dass leserseitig insbesondere die Geschwindigkeit und Sicherheit der Worterkennung, die Arbeitsgedächtniskapazität und das zur Verfügung stehende Vorwissen für das Zustandekommen interindividueller Unterschiede im Leseverständnis zentral sind (Schneider, 2008). Aufgrund der Vielzahl bereits untersuchter Faktoren und Befunde ist es nicht möglich, diese vollständig und im Detail darzustellen, vielmehr gibt das folgende Kapitel einen kleinen Einblick in entsprechende Forschungsergebnisse.

7.1 Leserseitige Einflüsse

Auf Seiten des Lesers wurden in der Forschung neben Wahrnehmungsprozessen kognitive Faktoren (z. B. Wortschatz, Arbeitsgedächtnis, Vorwissen), metakognitive Faktoren (z. B. Strategiewissen, Überwachungs- und Regulationsprozesse), motivationale Faktoren (z. B. Lesefreude und Interesse) und biologische Faktoren (z. B. Neurophysiologie und genetische Veranlagung) betrachtet.

Wahrnehmungsprozesse. Unterschiede in Wahrnehmungsprozessen (z. B. Blickbewegungen, Wahrnehmungsspanne, Buchstabenerkennung oder Diskriminationsleistung) können inzwischen als Ursache für interindividuelle Differenzen in der Lesekompetenz weitgehend ausgeschlossen werden (vgl. z. B. Artelt et al., 2007; Schnotz & Dutke, 2004; Rosebrock & Nix, 2011).

Wortschatz. Thorndike (1973, S. 61f.) fand eine hohe Korrelation zwischen Wortschatzumfang und Leseverständnis. Auch in der Frankfurter Hauptschulstudie ($N = 527$) erwies sich der Wortschatz als bedeutendste Determinante und einziger zuverlässiger Prädiktor für flüssiges Lesen (Gold, 2009). Weitere Studien belegten zudem, dass Texte mit vielen unbekannten Wörtern schlechter verstanden werden (Artelt et al., 2007).

Einerseits wurde zudem mehrfach der Erfolg von Wortschatztrainings zur Verbesserung der Worterkennung und des Leseverständnisses nachgewiesen. Auch wenn in einzelnen Studien, z. B. der von Tuinman und Brady (1974), vorheriges Üben des Vokabulars das spätere Textverständnis bei Grundschülern nicht verbesserte, zeigten in der Mehrzahl der Studien vor allem die Vermittlung eines breiten, leicht zugänglichen und kontextunabhängigen Wissens über Wortbedeutungen sowie eine tiefe und vorwissensgestützte Verarbeitung positive Effekte (I. L. Beck, Perfetti & McKeown, 1982; I. Beck, McKeown & Omanson, 1987; Kameenui, Carnine & Freschi, 1982; McKeown, Beck, Omanson & Perfetti, 1983). Andererseits zeigte sich umgekehrt, dass eine hohe Lesekompetenz den Erwerb neuer Wörter begünstigen kann (Artelt et al., 2007; Schnotz & Dutke, 2004). Bei Sternberg und Powell (1983) korrelierte der Wortschatz in gleicher Höhe mit der Leseleistung wie die Fähigkeit, Wortbedeutungen aus dem Kontext zu erschließen. Insgesamt sind die Befunde zum Einfluss des Wortschatzes also wenig konsistent. Der Wortschatz scheint zwar beim Leseverständnis eine Rolle zu spielen, jedoch nicht die entscheidende. Zudem scheint eine bidirektionale Beziehung zwischen Wortschatz und Leseverständnis vorzuliegen.

Worterkennung. Auch die schnelle und korrekte Identifikation von Wörtern ist für die erfolgreiche Bewältigung schriftlicher Texte erforderlich (Landerl & Willburger, 2009). Schlechtes Leseverständnis scheint häufig in defizitärer Worterkennung bzw. langsamem Zugriff auf das mentale Lexikon begründet zu sein (vgl. Artelt et al., 2007). Die Schnelligkeit des Zugriffs auf das mentale Lexikon kovariiert u. a. mit der Lesegeschwindigkeit und dem Leseverständnis (Hunt, Lunneborg & Lewis, 1975; Jackson & McClelland, 1979).

Die Ergebnisse korrelativer Studien können allerdings nicht kausal interpretiert werden und bei Fleisher, Jenkins und Pany (2013) zeigte sich, dass schwache Leser der vierten und fünften Klasse durch Trainingsmaßnahmen in der Worterkennung diesbezüglich zu den guten Lesern aufschlossen, sich ihr Leseverständnis jedoch nicht verbesserte. Insgesamt scheint eine unzureichende Automatisierung der Worterkennung

nicht genug kognitive Kapazitäten für hierarchiehöhere Verarbeitungsprozesse zu lassen. Ein schnellerer Zugriff auf das mentale Lexikon scheint das Verständnis aber noch nicht sicherzustellen.

Arbeitsgedächtnis. Das verbale Arbeitsgedächtnis ist schon zu Beginn des Schriftspracherwerbs ein einflussreicher Faktor, da beim Erlesen von Wörtern Buchstabe für Buchstabe die zuerst gelesenen Buchstaben noch verfügbar sein müssen, um am Ende die korrekte Wortbedeutung abrufen zu können. Auch später müssen zur Herstellung lokaler und globaler Kohärenz viele Informationen integriert werden (vgl. Artelt et al., 2007; Schnotz & Dutke, 2004).

Bei Daneman und Carpenter (1980, 1983) konnten Personen mit einer größeren Arbeitsgedächtnisspanne Pronomialreferenzen leichter auflösen und mehrdeutige Wörter besser kontextangemessen interpretieren als Personen mit einer kleineren Gedächtnisspanne. Auch bei de Jonge und de Jong (1996) zeigten sich bei Viert-, Fünft- und Sechstklässlern signifikant positive Korrelationen zwischen den Ergebnissen eines standardisierten Leseverständnistests sowie einfachen und komplexen Maßen der Gedächtnisspanne. Bei Studierenden fanden sich noch deutlich höhere Zusammenhänge zwischen der Arbeitsgedächtnisleistung und der Leseverständnisleistung (vgl. Hacker & Osterland, 1995). Friedman und Miyake (2000) ließen ihre Probanden Texte lesen und forderten sie auf, während des Lesens Aussagen zu beurteilen, welche eine Bildung von Inferenzen erforderten. Sie fanden einen generellen Zusammenhang zwischen der Ausbildung eines Situationsmodells und Arbeitsgedächtnismaßen. Palladino, Cornoldi, De Beni und Pazzaglia (2001) untersuchten bei Personen mit höherem und geringerem Leseverständnis eine Arbeitsgedächtnisleistung, die als Updating bezeichnet wird und das gezielte Präsenthalten und die Anpassung von Textinformationen während des Lesens meint. Es zeigte sich ein deutlicher Zusammenhang zwischen der Updating-Leistung und dem Leseverständnis.

Es liegen jedoch auch Studien vor, die dafür sprechen, dass sich eine geringere Arbeitsgedächtniskapazität nicht zwangsläufig negativ auf die Informationsverarbeitung beim Lesen auswirken muss, wenn die Kapazität optimal genutzt und auf die wesentlichen Informationen fokussiert wird (King & Just, 1991).

Inhaltliches Vorwissen. Beim inhaltlichen Vorwissen wird vermutet, dass es die Inferenzbildung unterstützt, indem es zusätzliche Verknüpfungen ermöglicht (vgl. Schnotz & Dutke, 2004). Inhaltliches Vorwissen kann geringere Lesefertigkeiten teilweise kompensieren und bei gleichen Verarbeitungsfähigkeiten unterschiedliche Leistungen erklären (B. C. Adams, Bell & Perfetti, 1995; Schneider, Körkel & Weinert, 1989; J. F. Voss & Silfies, 1996). Ein Mindestmaß an Vorwissen ist für das Verständnis fast jeden Textes erforderlich, um notwendige Inferenzen zu bilden (Schnotz, 1994; Singer, Graesser & Trabasso, 1994). Der Aufbau einer umfangreichen Vorwissensbasis und die Lesekompetenz beeinflussen sich von Anfang an gegenseitig, da einerseits Vor-

wissen für aktiv-konstruktive, inferenzielle Prozesse benötigt wird und andererseits der Erwerb von Vorwissen aus schriftlichen Texten durch eine hohe Lesekompetenz unterstützt wird (Artelt et al., 2007).

Verschiedene Studien zeigten, dass die Textbasis leichter memorisiert werden kann, wenn besonders viel Vorwissen zu ihrer Generierung herangezogen wurde (Chiesi, Spilich & Voss, 1979; Spilich, Vesonder, Chiesi & Voss, 1979). Für den Aufbau eines Situationsmodells spielt die Aktivierung von inhaltlichem Vorwissen sogar eine noch größere Rolle. Dies zeigten z. B. Tardieu, Ehrlich und Gyselinck (1992), die den Probanden Texte zu Themen vorlegten, zu denen diese viel oder wenig Vorwissen hatten. Während des Lesens sollten die Probanden mehrfach MC-Fragen beantworten, die sich entweder auf die Textbasis bezogen und eine Paraphrasierung eines zuvor gelesenen Satzes darstellten oder sich auf das Situationsmodell bezogen, indem Inferenzen aus mehreren zuvor gelesenen Sätzen erforderlich waren. Bezüglich der Paraphrasierungsfragen zeigte sich kein Unterschied in der Antwortzeit zwischen Personen mit viel und Personen mit wenig Vorwissen, bezüglich der Fragen zum Situationsmodell waren jedoch die Personen mit viel Vorwissen jenen mit wenig Vorwissen überlegen. Letzteres wurde als Hinweis auf ein elaborierteres Situationsmodell von Personen mit viel Vorwissen interpretiert. Zudem scheint Vorwissen beim Aufbau eines Situationsmodells Defizite der Textbasis kompensieren zu können. (Dutke, 1993, 1994, 1996). In Bezug auf inhaltliches Vorwissen fallen die Befunde somit konsistent aus und zeigen einen positiven Zusammenhang zwischen dem Ausmaß an Vorwissen und dem Leseverständnis, wobei eine wechselseitige Beeinflussung stattfindet.

Wissen über Textgenres. Nicht nur inhaltliches Vorwissen spielt für das Leseverständnis eine Rolle, sondern auch Wissen über bzw. Erfahrungen mit verschiedenen Textgenres. Textgenres weisen in der Regel bestimmte vorgegebene oder charakteristische Strukturen auf, sodass das Wissen über das Genre eines vorliegenden Textes das Textverständnis im Sinne einer Vorstrukturierung erleichtern kann (Artelt et al., 2007). So kann relevantes vorhandenes Wissen voraktiviert und zum Schließen von Kohärenzlücken herangezogen werden, was zudem die Lesegeschwindigkeit erhöht. Vor allem für literarische Texte wird angenommen, dass Wissen über das vorliegende Textgenre besonders wichtig ist – sogar wichtiger als inhaltliches Vorwissen (Christmann & Groeben, 1999).

Lesestrategien und Metakognition. Eine weitere Wissensart, die für das Leseverständnis von Bedeutung ist, ist das Wissen über Lesestrategien (Artelt et al., 2007). Zudem sind metakognitive und metalinguistische Fähigkeiten wichtige Faktoren für das Leseverständnis (D. H. Rost & Buch, 2010; Wellman, 1983). Im Laufe der gesamten Schulzeit – also auch noch in der Sekundarstufe – entsteht aufgrund vielfältiger Lernerfahrungen ein umfangreiches Strategiewissen (Schneider, 1996; Artelt et al., 2007; Schlagmüller, Visé & Schneider, 2001). Selbst bei Kontrolle von Dekodierfähigkeit, ko-

gnitiver Grundfähigkeit, verbalem Selbstkonzept und Leseinteresse leistet dieses Strategiewissen noch einen deutlichen Beitrag zur Vorhersage der Leseleistung. Das Wissen über Lesestrategien erwies sich bei PISA 2000 nach der kognitiven Grundfähigkeit als zweitbester Prädiktor für die Leseleistung (Artelt, Schneider & Schiefele, 2002).

Bei den Lesestrategien kommt es jedoch nicht nur darauf an, sie zu kennen, sondern auch darauf, sie anzuwenden. Dass Lernstrategien und metakognitive Überwachung Einfluss auf das Textverstehen und das Lernen mit Texten haben, wurde bereits vielfach empirisch belegt (z. B. Alexander & Judy, 1988; Gold, 2007; Groeben, 1982; Körkel & Hasselhorn, 1987; Mandl & Ballstaedt, 1982; Pressley, Wood & Woloshyn, 1990). Vor allem für hierarchiehöhere Leseprozesse ist eine Anwendung von Lesestrategien bedeutsam (Artelt et al., 2007). Schüler mit schwachen Verständnisleistungen haben häufig Probleme, spontan das Verständnis zu kontrollieren, vor allem bei Texten, die widersprüchliche oder mehrdeutige Informationen enthalten (Schnotz & Dutke, 2004). Ihr Verständnismonitoring erfolgt oft nur innerhalb eines Satzes. Es werden kaum Strategien angewendet, um den Verstehensprozess über längere Abschnitte aufrechtzuerhalten. Gute Leser wenden dagegen während des Lesens Strategien an, wie z. B. Aktivierung von Vorwissen, Konzentration auf das Erfassen der Kernideen und Bewertung der Kernideen hinsichtlich Konsistenz (Artelt et al., 2007; Ciborowski, 1992).

Insgesamt haben die meisten Lernstrategien einen signifikant positiven Einfluss auf die Leseleistung (s. z. B. Anderson & Armbruster, 1984; Kardash & Almund, 1991; Kiewra, 1989; Kletzien, 1991; Pressley et al., 1990). Bisherige Befunde deuten darauf hin, dass die exekutive metakognitive Komponente (Planen, Überwachen und Steuern) besonders wichtig ist. D. H. Rost und Buch (2010) vermuten allerdings, dass die Vorhersagekraft der Metakognition stark gemindert oder sogar eliminiert wird, wenn man den Einfluss der (verbalen) Intelligenz herauspartialisiert.

Motivation und Interesse. Inzwischen konnte recht gut belegt werden, dass motivationale Faktoren – vermittelt über wichtige Merkmale des Leseverhaltens, vor allem die Lesemenge und die Anwendung von Lesestrategien – einen Einfluss auf die Lesekompetenz haben (vgl. Möller & Schiefele, 2004; Paris, Wasik & Turner, 1991; Baker & Wigfield, 1999; Guthrie, Wigfield, Metsala & Cox, 1999; Wigfield & Guthrie, 1997). Cipielewski und Stanovich (1992) fanden beispielsweise, dass viel lesende Schüler eine höhere Lesekompetenz aufweisen als ihre wenig lesenden, gleich intelligenten Altersgenossen. Einer Schätzung von Cunningham und Stanovich (1997) zufolge erklärt die Lesemenge knapp ein Viertel der Leseverständnissteigerung von der fünften zur zehnten Klasse.

Die Analyse der PISA 2000-Daten ergab, dass neben einigen anderen Faktoren auch die intrinsische Motivation bedeutsam zur Erklärung interindividueller Differenzen in der Leseleistung bei den Schülern beiträgt (Möller & Schiefele, 2004). Zudem zeigten Schüler aus sozial schwachen Familien mit einer hohen intrinsischen Lesemotivation bessere Leseleistungen als Schüler aus Familien mit hohem sozioökonomischen

Status (SÖS) und niedriger Lesemotivation (I. Kirsch et al., 2002, S. 129). Bei DESI wiesen zudem Schüler mit hohem Leseinteresse schon am Anfang der neunten Klasse eine höhere Lesekompetenz auf, sie verfügten über differenziertere Lesestrategien und setzten sich unter Zuhilfenahme unterschiedlicher Techniken mit Texten auseinander (DESI-Konsortium, 2006, S. 36). Auch ganze Klassen mit einem höheren durchschnittlichen Leseinteresse hatten bereits zu Beginn der neunten Klasse ein höheres Lesekompetenzniveau und wiesen über das Schuljahr hinweg einen größeren Anstieg auf als Klassen mit niedrigerem Interesse. Insgesamt zeigt sich also konsistent ein positiver Zusammenhang von Lesemotivation und Leseinteresse mit der Leseverständnisleistung.

Biologische Einflüsse. Neurophysiologische und genetische Einflüsse auf die Lesekompetenz werden mittels bildgebender, nicht invasiver Verfahren, mittels Postmortem-Studien sowie mittels Verwandtschaftsstudien und molekulargenetischer Analysen zu identifizieren versucht. Verschiedene *neurophysiologische und Postmortem-Studien* fanden Unterschiede zwischen Personen mit und ohne LRS in der Aktivität und der Zellorganisation in bestimmten Gehirnarealen (z. B. Rumsey et al., 1992, 1997; Galaburda, Sherman, Rosen, Aboitiz & Geschwind, 1985; Cattinelli, Borghese, Gallucci & Paulesu, 2013). Es handelt sich jedoch um Kovarianzanalysen, weshalb unklar ist, ob die beobachteten Auffälligkeiten bei Personen mit LRS Ursache oder Folge der Schriftsprachprobleme sind (H. Marx & Reinhold, 2010). Allerdings konnte in einer Studie von Costanzo, Menghini, Caltagirone, Oliveri und Vicari (2012) nichtinvasive magnetische Stimulation bestimmter Gehirnareale die Genauigkeit beim Pseudowortlesen verbessern.

Auf *Verwandtschaftsstudien* basierende Schätzungen des Anteils interindividueller Varianz bezüglich des Lesens und leserelevanter Prozesse, der durch genetische Ausstattung erklärt wird, liegen bei 40 % bis 60 % (Grigorenko, 2004). Die Höhe des erklärten Varianzanteils variiert abhängig vom Alter der Person (niedrigerer Anteil bei jüngeren Personen), von der Sprache und von der sozialen Umgebung (z. B. SÖS, Schulsystem). Als höher wird der genetische Einfluss beim Auftreten von Lesestörungen bzw. Lernbehinderungen angesehen (Grigorenko, 2011). Heute gilt als weitgehend akzeptiert, dass bei LRS genetische Faktoren ursächlich beteiligt sind.

Molekularstudien verwenden Genmaterial und Ergebnisse von Leseaufgaben von starken und schwachen Lesern und prüfen, welche Gene mit schwachen Leseleistungen korrelieren. Inzwischen wurden etwa 20 (die Anzahl variiert abhängig von der LRS-Definition) möglicherweise relevante Genomregionen sowie mindestens sieben sogenannte „Kandidatengene“ (DYX1C1, KIAA0319, DCDC2, ROBO1, MRPL2, C2ORF3 und CYP19A1) innerhalb dieser Genomregionen identifiziert, die signifikant mit LRS in Zusammenhang stehen (Grigorenko, 2011; Anthoni et al., 2012). Jedoch ist bisher keine Genomregion und kein Kandidatengen von Genforschern allgemein akzeptiert oder abgelehnt. Die Forschung in diesem Bereich befindet sich noch in den Kinderschuhen (Grigorenko, 2011). Insgesamt ist jedoch unumstritten, dass – zumin-

dest beim Vorliegen einer umschriebenen Lesestörung – biologische Faktoren bei der Lesekompetenz eine Rolle spielen.

7.2 Einflüsse des sozialen Umfeldes

Wie in den vergangenen Kapiteln bereits mehrfach anklang, hat das soziale Umfeld großen Einfluss auf die individuelle Ausprägung der Lesekompetenz. Dabei spielen verschiedene Faktoren eine Rolle, von denen im Folgenden der familiäre Hintergrund und das schulische Umfeld betrachtet werden.

Familiärer Hintergrund. Schwache Leser kommen häufig aus Familien mit niedrigem SÖS und/oder Familien mit Migrationshintergrund und erfahren im familiären Umfeld wenig Leseförderung (Stanat & Schneider, 2004; Artelt et al., 2007; Schneider & Pressley, 1989; van Kraayenoord & Schneider, 1999; Bos et al., 2003). Studien zeigten einen Zusammenhang zwischen LRS und mütterlicher Schulbildung, elterlichem Bildungsabschluss, förderlicher Eltern-Kind-Interaktionen sowie übermäßigem Fernsehkonsum (z. B. Klicpera & Gasteiger Klicpera, 1993; H. Marx & Reinhold, 2010).

Bei PISA 2009 konnten 13 % der Variabilität der Leseleistung der 15-Jährigen durch Statusmerkmale (z. B. Bildungsniveau, Erwerbstätigkeit und Beruf der Eltern, Familienzusammensetzung) erklärt werden (Ehmke & Jude, 2010). Zudem zeigte sich ein negativer Zusammenhang der Lesekompetenz mit dem Vorhandensein eines Fernsehers bzw. einer Spielekonsole im eigenen Zimmer und ein positiver Zusammenhang mit dem Vorhandensein literarischer Texte und Hörbücher. Schüler mit eigenem Fernseher bzw. eigener Spielekonsole waren unter den schwachen Lesern mit 21 % überrepräsentiert, während Schüler ohne eigenen Fernseher bzw. eigene Spielekonsole mit 11 % unterrepräsentiert waren (Naumann et al., 2010). Während die PISA-Daten aufgrund des Querschnittsdesigns keine Kausalschlüsse zulassen, gibt es Studien, die experimentell bzw. im Längsschnittdesign zeigten, dass ein stark ausgeprägter Fernsehkonsum ebenso wie das Spielen am Computer negative Effekte auf die Lesekompetenzentwicklung haben können (Ennemoser & Schneider, 2007; Weis & Cerankosky, 2010). Ennemoser und Schneider (2007) fanden Hinweise darauf, dass bei den Auswirkungen des Fernsehkonsums verschiedene Wirkfaktoren zusammenspielen und akkumulieren. Ihre Ergebnisse deuten insbesondere darauf hin, dass Fernsehkonsum die Lesekompetenz beeinträchtigt, indem er Leseaktivitäten verdrängt, sowie, dass das Vergnügen, das Fernsehen den Schülern bereitet, ihre Bereitschaft, sich beim Lesenlernen anzustrengen, verringert. Ungünstig schien sich der Fernsehkonsum insbesondere dann auszuwirken, wenn er ausschließlich der Unterhaltung diente und wenig ausdifferenziert war (vgl. Schneider, 2008). Auf der anderen Seite waren bei PISA 2009 Schüler aus Haushalten mit Hörbüchern, Büchern mit Gedichten und klassischer Literatur auf den untersten beiden Niveaustufen mit ca. 8 %, 9 % und 7 % unter-

repräsentiert (Naumann et al., 2010). Bei statistischer Kontrolle der Lesemotivation und des Lesestrategiewissens zeigte sich allerdings bei PISA 2009 kein Unterschied mehr in der Leseleistung zwischen Schülern aus bildungsnahen und bildungsfernen Elternhäusern (Artelt et al., 2010).

Schüler mit Migrationshintergrund gehören bezüglich des Schriftspracherwerbs und der Entwicklung von Lesekompetenz häufig zur Risikogruppe (Artelt et al., 2007; Baumert & Schümer, 2001; Bos et al., 2003). Bereits am Ende der Grundschulzeit schneiden Kinder mit Migrationshintergrund in der Lesekompetenz deutlich schlechter ab als Kinder ohne Migrationshintergrund. Dies bleibt auch im weiteren Verlauf der Schulzeit so, wie die PISA-Ergebnisse zeigen (vgl. Kap. 6). Da die meisten Kinder mit Migrationshintergrund bilingual aufwachsen, werden als Erklärung Aspekte der Bilingualität – z. B. Interferenzen der beiden Sprachen, geringerer Wortschatz in beiden Sprachen und geringere Vertrautheit mit linguistischen Strukturen – diskutiert. Der Migrationsstatus ist zudem häufig mit einem ungünstigeren sozialen Umfeld – vor allem im Hinblick auf Lese- und Schreibförderung – verknüpft. Vermittelt über die Menge an Lerngelegenheiten wirken sich auch das Alter bei der Zuwanderung, die Aufenthaltsdauer in Deutschland und die in der Familie gesprochene Sprache auf die Lesekompetenz aus (Alba & Nee, 1997; Baumert & Schümer, 2002; H. Esser, 1990; Stanat, 2003; Stanat & Schneider, 2004). Bei den meisten Erklärungsansätzen handelt es sich um Vermutungen, für die es erste Hinweise gibt; zur Untermauerung der Hypothesen und zum Verständnis genauer Wirkprozesse ist weitere Forschung nötig.

DESI fand einen deutlichen Zusammenhang zwischen SÖS und Erstsprache, wobei Schüler mit Deutsch als Erstsprache aus Familien mit einem deutlich höheren SÖS stammen als solche mit nichtdeutscher Erstsprache; der Status der Familien mit bilingualen Kindern liegt dazwischen (DESI-Konsortium, 2006, S. 22f.). Aus diesem Grund war der Einfluss des SÖS bei der Untersuchung des Einflusses des Migrationshintergrundes stets zu kontrollieren. Bezüglich der Lesekompetenz zeigte sich ein deutlicher Vorsprung der Schüler mit Deutsch als Erstsprache gegenüber den beiden anderen Gruppen, wobei Schüler mit nichtdeutscher Erstsprache die schwächsten Leistungen zeigten. Der Leistungszuwachs im Laufe der neunten Klasse war in allen drei Gruppen gleich hoch. Bei Kontrolle der mit Erstsprache in bedeutsamem Zusammenhang stehenden Variablen Schulart, SÖS, Geschlecht und kognitive Grundfähigkeit lagen Schüler mit Migrationshintergrund weiterhin zurück. Besonderer Förderbedarf schien bei nichtdeutschsprachigen Schülern im Wortschatz zu bestehen. Auch bei PISA 2009 zeigte sich, dass Schüler mit Migrationshintergrund in Deutschland eher aus sozial schwachen Familien stammen als Schüler ohne Migrationshintergrund (Stanat et al., 2010). Zudem zeigte sich, dass in fast allen Staaten unter den Schülern mit Migrationshintergrund (erste und zweite Generation) Schüler mit schwachen und sehr schwachen Lesekompetenzen deutlich überrepräsentiert sind, wobei Deutschland mit einem Anteil von ca. 10 % schwacher und sehr schwacher Leser unter den Migranten im Durchschnittsbereich liegt (Naumann et al., 2010). Im Bereich des sehr guten

Lesens (Niveaustufe V und VI) lagen jedoch nur etwa 2.5 % der Schüler mit Migrationshintergrund.

Stanat und Schneider (2004) verglichen anhand der PISA 2000-Daten die Prädiktoren für schwache Leseleistungen bei Schülern mit vs. ohne Migrationshintergrund und fanden keine bedeutsamen Unterschiede. Vielmehr scheint es sich bei den schwachen Lesern um eine recht homogene Gruppe zu handeln. Einzig in Bezug auf den Geschlechterunterschied zeigte sich ein Unterschied zwischen Schülern mit und ohne Migrationshintergrund. Zwar wurde in beiden Gruppen ein Geschlechterunterschied zugunsten der Mädchen gefunden, jedoch ließ er sich bei Schülern ohne Migrationshintergrund durch geringeres Interesse der Jungen erklären, während dies bei Schülern mit Migrationshintergrund nicht der Fall war. Auch die aufgrund häufig vorliegender kumulativer Misserfolgserlebnisse im Laufe der Schulzeit vermuteten Motivationsdefizite bei Schülern mit Migrationshintergrund im Vergleich zu Schülern ohne Migrationshintergrund konnten nicht bestätigt werden – tendenziell zeigte sich sogar das Gegenteil (Kao & Tienda, 1995; Stanat & Schneider, 2004). Bezüglich der phonologischen Bewusstheit wird davon ausgegangen, dass bilinguale Kinder nicht benachteiligt sind. Es liegt eher das Gegenteil nahe. Zudem profitieren Kinder mit Migrationshintergrund auch von Fördermaßnahmen zur phonologischen Bewusstheit (Schneider & Küspert, 2003; Penner, 2003; Weber, Marx & Schneider, 2007).

Bei differenzierterer Betrachtung zeigten sich bezüglich der Leseflüssigkeit auf Wortniveau keine Unterschiede zwischen Grundschülern mit und ohne Migrationshintergrund, teilweise auch nicht bei kurzen Texten (2-3 Sätze) (Limbird & Stanat, 2006; A. E. Marx & Stanat, 2011). Bezüglich des Hörverstehens wurden jedoch Unterschiede gefunden. Als Erklärung für das geringer ausgeprägte Hörverstehen stehen der geringere Wortschatz, geringere grammatikalische Fähigkeiten, geringere Automatisierung aufgrund reduzierten Kontakts mit der deutschen Sprache sowie geringeres bereichsspezifisches Vorwissen aufgrund eines anderen soziokulturellen Hintergrunds zur Diskussion (z. B. Artelt et al., 2007; Limbird, 2007).

Schulische Einflüsse. In Bezug auf schulische Faktoren werden z. B. Unterrichtsmethoden, die Lehrerpersönlichkeit und die Lehrer-Schüler-Interaktion untersucht, wobei die diesbezüglichen empirischen Belege für zuverlässige Aussagen bislang noch unzureichend sind (H. Marx & Reinhold, 2010). Auch zum Einfluss der Schulart liegen erste Ergebnisse vor.

In einer Studie von Morrow (1996, S. 74) führte ein ansprechendes Programm, das die Erhöhung der Lesezeit und Gelegenheiten zu gemeinsamem Lesen und Schreiben im Unterricht beinhaltete, im Vergleich zu einer Kontrollgruppe zu besseren Leseverständnisleistungen. Dies deutet darauf hin, dass nicht nur die Lesemenge in der Freizeit, sondern auch der Umfang schulischer Leseaktivitäten die Lesekompetenz beeinflussen (vgl. auch Elley, 1992; Metsala & Ehri, 1998). DESI untersuchte unter anderem den Zusammenhang zwischen der Leseförderung im Deutschunterricht

und der Lesekompetenz sowie motivationalen und affektiven Komponenten (DESI-Konsortium, 2006, S. 36). Dabei zeigte sich, dass lesespezifische Lerngelegenheiten im Deutschunterricht die Lesemotivation fördern. Dies ist vor allem dann der Fall, wenn didaktische Materialien die Auseinandersetzung mit unterschiedlichen Texten anregen. Negative Auswirkungen auf die Lesemotivation der Neuntklässler hatte nur ein aus Schülersicht zu hohes und überforderndes Unterrichtstempo. Weiter bestand bei DESI kein Zusammenhang zwischen der reinen Vielfalt an verwendeten Methoden und Textsorten und dem Kompetenzanstieg der Neuntklässler über das Schuljahr hinweg. Lediglich die Arbeit mit Prosatexten hatte Einfluss auf die Lesemotivation und die Lesekompetenz (DESI-Konsortium, 2006, S. 35f.). Eine wichtige Rolle spielte es zudem, ob die Schüler das Gefühl hatten, ihrer Lehrkraft seien sprachliche Kompetenzen wichtig. Diejenigen Klassen, die der Meinung waren, sprachbezogene Fähigkeiten seien im Unterricht sehr wichtig, zeigten bereits zu Beginn der neunten Klasse bessere Leistungen im DESI-Test sowie einen größeren Kompetenzzuwachs im Verlauf des Schuljahres. Insgesamt erwies sich bei DESI ein anspruchsvoller, aber verständlicher Unterricht mit wenig Tempodruck und mit methodischer Vielfalt als förderlich für das Leseverständnis (Philipp, 2011a).

Retelsdorf und Möller (2011) untersuchten den Einfluss der Schulart auf die Entwicklung der Leseleistung in der Sekundarstufe. Hier wurden anhand von Propensity Score Matching vergleichbare Schülerpaare gebildet und Hinweise darauf gefunden, dass die Schulart keinen Einfluss auf die Verständnisleistung zu haben scheint – diesbezüglich wurde keine Öffnung der Leistungsschere zwischen Gymnasiasten und Nicht-Gymnasiasten gefunden – während die Schulart sich auf die Dekodiergeschwindigkeit auszuwirken scheint, da sich hier eine Öffnung der Leistungsschere zeigte.

7.3 Textseitige Einflüsse

Nicht nur der Leser selbst oder sein soziales Umfeld, sondern auch der Text ist eine Quelle für inter- und intraindividuelle Differenzen im Leseverständnis. Dabei geht es weniger um die generelle Fähigkeit einer Person, sinnentnehmend zu lesen, sondern vielmehr darum, einen bestimmten Text zu verstehen. Die Textverständlichkeitsforschung hat bereits vielfach gezeigt, dass bestimmte Textmerkmale die Lesbarkeit und Verständlichkeit fördern.

Textoberfläche. Im Hinblick auf die Textoberfläche lassen sich als Einflussfaktoren die typographische Gestaltung und sprachformale Merkmale anführen (D. H. Rost & Buch, 2010). Bezüglich der *typographischen Gestaltung* erwiesen sich ein Mindestzeilenabstand, eine optimale Zeilenlänge sowie eine Mindestgröße für die Schrift als förderlich für den Lesefluss. Zur Berurteilung der Textverständlichkeit auf der Grundlage *sprachformaler Merkmale* gibt es sogenannte „Lesbarkeitsindizes“, die lexikalische und syntaktische Aspekte (z. B. mittlere Wort- und Satzlänge) heranziehen, um die

Verständlichkeit eines Textes daraus vorherzusagen. Semantische Aspekte werden dabei jedoch nicht berücksichtigt. Als Beispiele seien der „Flesch-Index“ (vgl. Bachmann, 2009) und der sogenannte „LIX“ (vgl. A. Lenhard & Lenhard, 2011) genannt. Der Flesch-Index wurde für den englischen Sprachraum entwickelt, ist jedoch auch im deutschen Sprachraum geläufig. Er geht davon aus, dass kurze Wörter und kurze Sätze üblicherweise leichter verständlich sind als lange Wörter und lange Sätze, wobei er die Wortlänge schwerer gewichtet als die Satzlänge. Der LIX ist demgegenüber ein an den deutschen Sprachraum angepasster Lesbarkeitsindex. Er resultiert aus der Summe der durchschnittlichen Satzlänge eines Textes und des prozentualen Anteils langer Wörter, d. h. Wörter mit mehr als sechs Buchstaben (vgl. A. Lenhard & Lenhard, 2011).

Inhaltliche Gestaltung. Über die Merkmale der Textoberfläche hinaus wurden faktorenanalytisch vier Dimensionen extrahiert, die sich unabhängig von der Thematik des Textes auf das Verstehen und Behalten des Inhaltes auswirken (vgl D. H. Rost & Buch, 2010): Einfachheit der Formulierung (z. B. Geläufigkeit der Wörter, Veranschaulichung abstrakter Inhalte, konkrete Ausdrucksweise), Gliederung des Textaufbaus (äußere Gliederung, logische Folgerichtigkeit, Übersichtlichkeit), Kürze und Prägnanz (Konzentration und verdichtete Darstellung des Wesentlichen) sowie zusätzliche stilistische Stimulanz (z. B. wörtliche Rede, Fragen an den Leser, Alltagsbeispiele). Einfachheit und Gliederung sollten für eine gute Verständlichkeit hoch ausgeprägt sein, Kürze und Prägnanz sowie stilistische Stimulanz sollten dagegen nur in mittlerem Ausmaß gegeben sein.

Insbesondere für den Aufbau eines Situationsmodells erwiesen sich eine kohärente Inhaltsorganisation (z. B. Hinweise auf Zusammenhänge, Kohäsionsmittel, typographische Markierungen), hierarchisch-sequenzielles Anordnen von Textinhalten (vom Abstrakten zum Detail) und die Aktivierung von Vorwissensbeständen (z. B. Advance Organizer) als förderlich, wobei die Textgestaltung stets Vorwissen und Erwartungen der Leser berücksichtigen sollte (Christmann & Groeben, 1996, 1999; Goldman & Rakestraw, 2000).

Textgenres. Es konnte bereits empirisch belegt werden, dass für verschiedene Textgenres bestimmte Schemata existieren – z. B. bei Erzähltexten sogenannte „Geschichtengrammatiken“ – die die Prinzipien von Aufbau und Struktur des jeweiligen Textgenres enthalten und das Verständnis erleichtern. Aber auch andere Texte, z. B. wissenschaftliche Aufsätze, Gebrauchsanweisungen oder Protokolle weisen eine immer gleiche Struktur auf, die das Lesen und Verstehen im Sinne einer Vorstrukturierung erleichtern. Mehrere Experimente von Zwaan (1993, S. 38ff.) z. B. zeigten, dass alleine eine Vorinformation über das Genre eines Textes (es handle sich dabei um einen literarischen Erzähltext vs. einen Zeitungsartikel) zur Anwendung unterschiedlicher Lesestrategien sowie schließlich unterschiedlichen mentalen Repräsentationen mit unterschiedlichem Informationsgehalt führte.

Grundsätzlich werden in der Literatur zwei große Textgenres unterschieden, wobei das eine Textgenre – expositorische Texte – typischerweise der Informationsvermittlung und das andere Textgenre – narrative Texte – der Unterhaltung dient. Sowohl narrative als auch expositorische Texte lassen sich noch weiter untergliedern, worauf hier jedoch nicht näher eingegangen werden kann; es sei diesbezüglich z. B. auf Rosebrock und Nix (2011) sowie Christmann und Groeben (2002) verwiesen. Die verschiedenen Genres lassen sich nicht immer exakt voneinander abgrenzen (Rosebrock & Nix, 2011, S. 76). Zudem sind narrative und expositorische Texte nicht per se unterschiedlich schwer zu verstehen, vielmehr gibt es bei beiden Genres verschiedene Schwierigkeitsstufen.

Narrative Texte sind häufig näher am realen Leben des Lesers, was dazu führt, dass die Verarbeitung von Erzählungen von den Lebenserfahrungen des Lesers beeinflusst wird (Adam-Schwebe et al., 2009). Daher werden beim Lesen narrativer Texte ähnliche kognitive Prozesse aktiviert wie bei der Verarbeitung realer Situationen. Die häufig mehrdeutigen Geschichten führen so zur Konstruktion komplexer, mehrschichtiger Situationsmodelle (Graesser et al., 1994; W. Kintsch, 1994; E. Kintsch & Kintsch, 1997). *Expositorische Texte* hingegen sind eher kontextunabhängig und dienen der Information über Konzepte, Sachverhalte oder technisch-wissenschaftliche Zusammenhänge (Adam-Schwebe et al., 2009). Entsprechend erwartet der Leser Objektivität, Klarheit und Eindeutigkeit. Häufig ist beim Lesen expositorischer Texte zudem wenig Vorwissen vorhanden. Daher werden selten multiple Schlussfolgerungen gezogen und eher einschichtige Situationsmodelle aufgebaut (Graesser et al., 1994; E. Kintsch & Kintsch, 1997). Die verschiedenen Textgenres stellen also unterschiedliche Anforderungen an den Leser und machen unterschiedliche Verarbeitungsschritte erforderlich.

7.4 Interaktionen

Bei der Betrachtung der Einflussfaktoren ist stets zu berücksichtigen, dass diese auf vielfältige Weise zusammenhängen und wechselwirken und die Wirkrichtung z. B. aufgrund querschnittlicher, nicht-experimenteller Designs nicht immer eindeutig bestimmt werden kann. Einige Wechselwirkungen wurden bereits im Verlauf dieses Kapitels angesprochen. Im Folgenden werden noch weitere Beispiele gegeben.

Zunächst wird die Interaktion von Lesermerkmalen und Textmerkmalen verdeutlicht. So ist die Bedeutung inhaltlichen Vorwissens z. B. abhängig von der Struktur, Schwierigkeit und Länge sowie von der Art eines Textes (Schiefele, 1996, S. 118). Bei anspruchsvollen und/oder wenig kohärenten Texten ist es von größerer Bedeutung als bei weniger anspruchsvollen und kohärenten Texten. Bei umfangreichem Vorwissen kann ein weniger kohärenter Text aktiv-konstruktive und inferenzielle Verarbeitung anregen und zum Aufbau eines besseren Situationsmodells führen, während das Feh-

len kognitiver Anreize zur Unterforderung führen und die Lesemotivation dämpfen kann (Christmann & Groeben, 2002; Groeben & Christmann, 1989; Groeben, 1978; McNamara, Kintsch, Songer & Kintsch, 1996). Außerdem sind abhängig vom Textgenre verschiedene Vorwissensarten von unterschiedlich großer Bedeutung. Beispielsweise spielt beim Verständnis von expositorischen Texten hauptsächlich inhaltliches Vorwissen eine Rolle, während dies bei literarischen Texten seltener nötig ist, dafür ist bei literarischen Texten in der Regel Textsortenwissen wichtiger (Rosebrock & Nix, 2011; Christmann & Groeben, 1999). Zudem sind für das Verständnis literarischer Texte affektive Beteiligung und Empathiefähigkeit von größerer Bedeutung als für das Verständnis expositorischer Texte.

Die Interaktion von biologischen und psychosozialen Faktoren wird bei der Betrachtung der Erklärungsansätze für die in vielen Studien gefundenen Geschlechterunterschiede (s. Kap. 6) besonders deutlich. Hier integriert z. B. das biopsychosoziale Modell von Halpern und Kollegen verschiedene Ansätze (Halpern & LaMay, 2000; Halpern & Tan, 2001): Biologische Befunde zu den Effekten funktionaler Lateralisierung der Gehirnhälften und zum Einfluss der Geschlechtshormone sprechen für leichte biologisch begründete Vorteile des weiblichen Geschlechts beim Erwerb verbaler Kompetenzen. Es handelt sich dabei jedoch um nicht deterministische Tendenzen, die durch gezielte Förderung ausgeglichen werden können. Die Umwelt bestärkt jedoch tendenziell die biologischen Veranlagungen, indem sie diese aufgreift, eine entsprechende Erwartungshaltung einnimmt und somit auf den weiteren Kompetenzerwerb einwirkt. Die erworbene Kompetenz wirkt schließlich wieder auf physiologische Funktionen zurück und verstärkt veranlagte Unterschiede weiter. Bei PISA 2000 konnten die Geschlechterunterschiede fast vollständig durch Motivationsunterschiede erklärt werden; bei gleichem Interesse zeigte sich kein Unterschied in der Lesekompetenz zwischen den Geschlechtern (Stanat & Kunter, 2001; Klieme & Stanat, 2002). Weiter gibt es Hypothesen, dass z. B. im Deutschunterricht vorwiegend Themen behandelt werden, die für Mädchen interessanter sind als für Jungen. Zudem könnte die große Mehrheit des weiblichen Geschlechts innerhalb der Lehrerkollegien in Sprachfächern dazu führen, dass für Jungen männliche Lesevorbilder fehlen (W. Lenhard, 2013, S. 61).

7.5 Fazit

Obwohl im vergangenen Kapitel jeweils nur ausgewählte Beispiele der leserseitigen, umgebungsbezogenen und textseitigen Faktoren sowie ihrer komplexen Interaktionen erläutert wurden, wurde die Vielfalt der Einflüsse und somit erneut die Vielschichtigkeit der Leseprozesse deutlich. Sollten sich im Rahmen einer Leseverständnisdiagnostik Hinweise auf bedeutsame Defizite ergeben, sollte daher eine umfassende Abklärung der genauen Ursachen folgen, wobei in erster Linie die schülerseitigen Aspekte zu

betrachten sind. Möglicherweise sollte aber darüber hinaus auch das soziale Umfeld herangezogen werden.

Aus einer Testdiagnostik ergeben sich idealerweise bereits Hinweise darauf, auf welcher Ebene des Leseverständnisses Defizite vorliegen, was die detaillierte Ursachensuche deutlich erleichtert. Weiter sollten sich aus der Diagnostik und den daraufhin ermittelten Ursachen Ansatzpunkte für adäquate Fördermaßnahmen ergeben. Zahlreiche Hinweise darauf, dass selbst in der Sekundarstufe schulische Maßnahmen noch zur Verbesserung der Leseleistung beitragen können, fanden sich in bereits genannten Studien (vgl. z. B. die im vorliegenden Kapitel dargestellten DESI-Befunde zu schulischen Einflüssen; DESI-Konsortium, 2006). Mit weiteren schulischen Fördermaßnahmen beschäftigt sich das folgende Kapitel.

Kapitel 8

Förderung von Leseverständnis

Während in der Grundschule die gezielte Leseförderung fester Bestandteil des Unterrichts ist, rückt dieser Aspekt in der Sekundarstufe zugunsten der Literaturarbeit als Fachunterricht in den Hintergrund (Rosebrock & Nix, 2011, S. 1). Zahlreiche Befunde zeigten jedoch, dass auch in der Sekundarstufe eine gezielte Förderung nötig ist und erfolgreich eingesetzt werden kann – wenn auch mit zunehmendem Alter der Schüler die Fördereffekte abzunehmen scheinen (vgl. Schneider, 2008).

An welchen Punkten entsprechende Fördermaßnahmen im Einzelnen ansetzen können, erschließt sich aus der Betrachtung der am Lesen beteiligten Prozesse, des Lesekompetenzerwerbs und der im vergangenen Kapitel dargestellten Einflussfaktoren auf die Lesekompetenz. Bei einigen Einflussfaktoren (z. B. Motivation, Selbstregulation, Wissen über Textmerkmale) ist es jedoch leichter als bei anderen (z. B. Arbeitsgedächtniskapazität), intervenierend einzugreifen (vgl. Streblow, 2004; Gold, 2007; Rühl & Souvignier, 2006; Mähler & Hasselhorn, 2000). Rosebrock und Nix (2011, S. 3) weisen weiter darauf hin, dass die Leseförderung an das jeweils vorliegende Defizit angepasst werden sollte, denn nicht alle Methoden sind gleichermaßen zur Lösung aller Leseprobleme geeignet. Während bei manchen Schülern auch noch in den mittleren und höheren Klassenstufen eine Förderung basaler Lesekompetenzen nötig ist, empfiehlt es sich in der Sekundarstufe generell, Strategien zu vermitteln und einzuüben (vgl. W. Lenhard, 2013, S. 133). Aber auch eine Vermittlung von Textformatwissen, von bereichsspezifischem Vorwissen und von Wortschatz sowie eine Förderung der Lesemotivation ist in der Sekundarstufe wichtig. Im Folgenden wird dargestellt, welche Maßnahmen empfohlen werden in Abhängigkeit davon, welche Komponente des Leseprozesses von einem Defizit betroffen ist. Beispiele von zur Verfügung stehenden Förderprogrammen werden genannt. Für detailliertere Informationen wird jedoch auf die aktuellen Werke von Rosebrock und Nix (2011), W. Lenhard (2013) sowie Philipp und Schilcher (2012) verwiesen.

8.1 Wortschatz

Zwischen Wortschatz und Lesekompetenz besteht ein bidirektionaler Zusammenhang (vgl. Kap. 6). Für Schüler, die nur einen geringen Wortschatz aufweisen – häufig ist das bei Schülern mit Migrationshintergrund der Fall (vgl. z. B. DESI-Konsortium, 2006, S.

25) – empfiehlt es sich daher Wortschatztrainings durchzuführen, die eine tiefe Verarbeitung der Bedeutung wichtiger Wörter beinhalten. Zusätzlich können im Rahmen von Lesestrategietrainings Strategien zum Herauslesen der Wortbedeutung aus dem Kontext vermittelt werden (Badel & Valtin, 2005). Die Wortschatzarbeit sollte generell im Kontext erfolgen, da Wörter und Sätze häufig erst im Kontext ihre Bedeutung erhalten (Spinner, 2006). Zudem ist es nicht das Ziel, einen abstrakten Wortschatz aufzubauen, sondern die Vernetzung der Wörter im mentalen Lexikon anzuregen. Empfehlungen hierfür beschreibt z. B. Kühn (2007).

8.2 Basale Lesekompetenzen

Für Schüler mit Defiziten im Bereich der basalen Lesekompetenzen werden vor allem Übungen zur Förderung der phonologischen Informationsverarbeitung, Dekodierübungen und Lautleseverfahren empfohlen (z. B. Artelt et al., 2007; Rosebrock & Nix, 2011; Gasteiger Klicpera & Klicpera, 2004). Es ist dabei zu berücksichtigen, dass Maßnahmen zur Förderung der Lesemotivation bei schwachen Lesern auch zu gegenteiligen Effekten führen können, wenn die geweckte Erwartung an ein positives Leseerlebnis aufgrund mangelnder Lesefertigkeit enttäuscht wird (Rosebrock & Nix, 2011, S. 93f.).

Phonologische Informationsverarbeitung. Bei LRS liegen in der Regel im Bereich der phonologischen Informationsverarbeitung charakteristische Defizite vor (Artelt et al., 2007). Entsprechend haben sich Trainingsmaßnahmen zur Phonemanalyse und Phonemsynthese bewährt, da es sich dabei um grundlegende Fähigkeiten handelt, die das Erlernen des alphabetischen Codes erleichtern. Für den Altersbereich ab der fünften Klasse ist hierfür z. B. das Verfahren „Lautgetreue Lese-Rechtschreibförderung“ (Reuter-Liehr, 2006) verfügbar. Im Rahmen dieses Trainings wird versucht, die schriftsprachlichen Kompetenzen von Grund auf neu aufzubauen. Ein Bestandteil des Trainings ist ein Üben der Wortdurchgliederung. Das Verfahren wirkt sich nachweislich positiv auf die Rechtschreibleistung aus, für die Leseleistung sind bislang keine belastbaren Befunde bekannt (Reuter-Liehr, 1993; Unterberg, 2005). Weitere Anregungen, die auch für ältere Schüler geeignet sind, geben Gasteiger Klicpera und Klicpera (2004).

Dekodieren und Leseflüssigkeit. Zur Förderung der Dekodierfähigkeit bieten sich intensive Dekodierübungen auf Wortebene zur Automatisierung der Worterkennung und zum Aufbau eines umfangreichen Sichtwortschatzes an. Hierfür sind z. B. im „Kieler Leseaufbau“ (Dummer-Smoch & Hackethal, 2011) Anregungen sowie Übungen enthalten.

Darüber hinaus empfehlen Rosebrock und Nix (2011, S. 8, 36ff.) zur Verbesserung der Leseflüssigkeit die Förderung hierarchieniedriger Prozesse auf Wort-, Satz- und lo-

kaler Textebene mithilfe von Lautleseverfahren. Dabei lesen Schüler kurze Texte laut vor, wodurch die Worterkennung, die Verbindung von Wortfolgen im Satzzusammenhang und die Herstellung von Beziehungen zwischen einzelnen Sätzen gefördert werden soll. Für sechste Hauptschulklassen wurde hierfür das Verfahren „Leseflüssigkeit fördern“ (Rosebrock, Nix, Rieckmann & Gold, 2011) positiv evaluiert. Bei dieser Tandem-Lautlese-Methode lesen ein lesestärkerer und ein leseschwächerer Schüler einen Textabschnitt mehrfach simultan und der lesestärkere Schüler trainiert den leseschwächeren Schüler im Sinne eines Trainers als Modell und Unterstützer praktisch wie einen Sportler.

8.3 Textverständnis

Für Schüler, die zwar über basale Lesekompetenz verfügen, jedoch Probleme haben, den Bedeutungsgehalt eines Textes zu rekonstruieren, bieten sich Maßnahmen zur Motivationsförderung, zum Wissen über und der Anwendung von Lesestrategien, zum Wissen über Textgenres und genrespezifische Herangehensweisen sowie zu thematischem Wissen an.

Motivation. Maßnahmen zur Steigerung der Lesemotivation gehen davon aus, dass eine höhere Motivation sich positiv auf die Lesehäufigkeit auswirkt und somit indirekt die Lesekompetenz erhöht wird (Artelt et al., 2007). Lesen soll als stimulierende, genussvolle und lohnende Freizeitaktivität entdeckt werden (Rosebrock & Nix, 2011, S. 92).

Im Rahmen der Motivationsförderung kann auf Individualebene z. B. bei der Verarbeitung von Leistungsrückmeldungen angesetzt werden, wobei die Leistungsattributionen und die Wahl der Referenzpersonen genauer betrachtet und verändert werden können, um negative Auswirkungen auf das Selbstkonzept zu vermeiden (Streblow, 2004). Wenn Schüler sich als kompetente Leser erleben, steigert das ihr Vertrauen in die eigenen Fähigkeiten und damit auch die Frustrationstoleranz, die Leseausdauer und letztendlich die Lesemotivation. Schiefele (2004, S. 138ff.) empfiehlt zur Förderung des Interesses neben der Kompetenzwahrnehmung auch die Selbstbestimmung, die soziale Einbindung und den Bedeutungsgehalt des Lerngegenstands in den Blick zu nehmen.

Für den Deutschunterricht bieten sich auf Klassenebene zahlreiche Verfahren und Veranstaltungen zur Leseanimation an, z. B. ein Einrichten einer Klassenbibliothek, oder ein Einsatz von Hörbüchern. Fächerübergreifend können Themen über Sachtexte erarbeitet und somit über das Sachinteresse die Lesemotivation geweckt werden. Auf Schulebene können z. B. Leseecken oder Lesecafés eingerichtet und Lesenächte veranstaltet werden. Schließlich kann Leseanimation über die Schule hinaus gehen, z. B. in Form einer Kooperation mit öffentlichen Büchereien oder der Organisation

von Autorenlesungen. Für zahlreiche weitere Möglichkeiten siehe z. B. Rosebrock und Nix (2011, S. 105ff.).

Strategien. Strategieförderprogramme empfehlen sich vor allem für Schüler, die Probleme haben, das Gelesene in einen Gesamtzusammenhang zu bringen und mit vorhandenem Wissen zu verknüpfen (Rosebrock & Nix, 2011, S. 62). Im Entwicklungsverlauf werden Strategien zunächst bereichsspezifisch erworben, ihr flexibler Einsatz erfolgt spontan recht spät und sollte daher unterstützt werden (Boekaerts, 1997). Durch eine Vermittlung und Einübung von Lern- und Lesestrategien soll explizit die Informationsverarbeitungskompetenz verbessert und die Fähigkeit zur Kohärenzbildung und zum Aufbau von Situationsmodellen unterstützt werden. Der effiziente Strategieeinsatz erfordert ein Repertoire an flexibel anwendbaren Strategien und Wissen über ihre Einsatzbedingungen, über Aufgabenanforderungen, über spezifische Merkmale von Lernmaterialien und über eigene Fähigkeiten, eigenes Wissen und eigene Einstellungen (Artelt et al., 2007). Weiterhin kann ein Lese- und Lernstrategietraining die Anwendung von Vorwissen beim Lesen fördern.

Lesestrategien haben den Vorteil, dass sie relativ gut für Interventionen zugänglich sind (z. B. Chi, Leeuw, Chiu & Lavancher, 1994; B. J. F. Meyer & Poon, 2001; Souvignier, Küppers & Gold, 2003; National Institute of Child Health and Human Development, 2000). Die Trainierbarkeit der Strategieanwendung im Umgang mit Textmaterial wurde bereits mehrfach belegt (vgl. Hasselhorn, 1983; Artelt, 2006). Neben der Vermittlung neuer Strategien und deren intensiver Einübung sollten suboptimale Strategien abgebaut werden (Artelt et al., 2007). Empirische Forschungsergebnisse deuten zudem darauf hin, dass die explizite Förderung von Strategiewissen und metakognitiver Kompetenz zum Aufbau eines positiven Fähigkeitsselbstbildes und leistungsförderlicher Attributionen führt, was sich wiederum nicht nur auf die Informationsverarbeitungskompetenz, sondern auch auf die Motivation auswirkt (Artelt et al., 2007).

Inzwischen stehen einige ausgearbeitete Trainingsprogramme zur Strategievermittlung zur Verfügung, die bereits für den Unterricht angepasst sind und deren Wirksamkeit in der Regel nachgewiesen ist (vgl. Rosebrock & Nix, 2011, S. 72). Für den deutschen Sprachraum am besten evaluiert sind wohl „Wir werden Textdetektive“ (Gold, Mokhlesgerami, Rühl, Schreblowski & Souvignier, 2004; Souvignier et al., 2003) und das speziell für sehr schwache Schüler und Schüler mit Lernbehinderung geeignete Programm „Wir werden Lesedetektive“ (Rühl & Souvignier, 2006; Antoniou & Souvignier, 2007). Eine etwas andere Herangehensweise verfolgt das computergestützte Trainingsprogramm „conText“ (W. Lenhard, Baier, Lenhard, Schneider & Hoffmann, 2013). Die Schüler fassen Texte zusammen und bekommen eine Rückmeldung über die Qualität dieser Zusammenfassung insbesondere im Hinblick auf die Inhaltsabdeckung, wodurch eine intensive Auseinandersetzung mit den Textinhalten angeregt wird, und somit auf implizitem Weg Strategien erworben werden. Das Programm

wurde für deutsche Hauptschüler evaluiert und zeigte positive Effekte für die Leseflüssigkeit und das Leseverständnis.

Textwissen und genrespezifische Förderung. Für den Umgang mit verschiedenen Textgenres, welche unterschiedliche Strategien und eine Aktivierung entsprechender Vorwissensbestände erfordern, empfehlen Artelt et al. (2007), Wissen über typische Textstrukturen sowie über spezifische Leseanforderungen, insbesondere des kritischen Lesens sowie des Reflektierens und Bewertens, zu vermitteln.

Zur Unterstützung der Sachtexterarbeitung kann z. B. die Generierung von Erwartungen an den Text und die Einordnung von Textaussagen gefördert werden, indem domänenspezifische Organisationsformen expositorischer Texte transparent gemacht werden (Rosebrock & Nix, 2011, S. 78). Von Dymock (2005) wird das Programm „CORE“ (Connect, Organize, Reflect, Extend) beschrieben, mit dessen Hilfe bei acht- bis zwölfjährigen Schülern im Unterricht das Erkennen von Textstrukturen beim Lesen von Sachtexten in mehreren Schritten gefördert werden kann. Die Unterstützung der Lektüre literarischer Texte sollte verschiedene Ebenen des Lesens ansprechen (Rosebrock & Nix, 2011, S. 121ff.). Dazu zählen hierarchiehöhere kognitive und nichtkognitive Prozesse, das Erkennen von Superstrukturen und Darstellungsstrategien sowie Wissen, Motivation, Beteiligung und Reflexion. Auch hier hat sich z. B. eine Vermittlung von Lesestrategien und Textsortenwissen als verstehensförderlich erwiesen. Der sowohl für schwache als auch für gute Leser gefundene Verstehenszuwachs fällt dabei für schwache Leser etwas größer aus (z. B. National Institute of Child Health and Human Development, 2000, S. 4ff.).

Bei schwachen Lesern, die Mühe haben, globale Kohärenz aufzubauen und begleitend zu differenzieren, können darüber hinaus z. B. Vorlesen oder eine Verwendung von Bildern und Medien, die inhaltliche Zusammenhänge darstellen, hilfreich sein. Szenisches Spielen von Textabschnitten kann zur detaillierten Imagination und zur affektiven Beteiligung anregen. Zur Förderung der Erfassung der Darstellungsintention empfehlen Rosebrock und Nix (2011, S. 124) Verfahren des handlungs- und produktionsorientierten Literaturunterrichts, indem die Schüler z. B. ein Gespräch mit einer literarischen Figur imaginieren und die Geschichte fortsetzen, wodurch der Aufbau von Situationsmodellen von jedem einzelnen Schüler gefordert und gefördert wird.

Vorwissen. Allgemein spielen für das Lesen neben dem bereits genannten Wortschatzwissen, dem metakognitiven Wissen und dem textsortenbezogenen Wissen auch thematisches Wissen und generelles Weltwissen sowie grammatisches Wissen eine Rolle (Köster, 2006; Schiefele, 1996). Im Hinblick auf das für das Leseverständnis besonders bedeutsame thematische Vorwissen empfehlen Artelt et al. (2007) systematisch inhaltlich relevantes Hintergrundwissen zu im Unterricht behandelten Texten zu vermitteln sowie bereits vorhandenes Wissen zu aktivieren und aufzufrischen. Zudem sollte beim Lesen in allen Schulfächern, die mit schriftlichen Texten arbeiten, systematisch eine Anwendung von Vorwissen geübt werden.

8.4 Kombination von Maßnahmen und Integration in den Unterricht

Laut W. Lenhard (2013, S. 146) kommt es insgesamt vor allem darauf an, Maßnahmen leistungs- und altersangemessen auszuwählen, eine Förderung langfristig anzulegen und verschiedene Elemente zu integrieren sowie verschiedene Sozialformen und unterschiedliche didaktische Materialien einzusetzen. Häufig wird eine Kombination aus Motivierung, Kompetenz- und Strategievermittlung empfohlen (z. B. von Artelt et al., 2007; Spinner, 2004). Im Hinblick auf schwache Leser, welche weder über die Kompetenz noch die Motivation zum verstehenden Lesen verfügen, erscheint insbesondere eine Kombination aus der Förderung basaler Prozesse und motivierenden Maßnahmen sinnvoll (vgl. z. B. Hurrelmann, 2006; Rosebrock & Nix, 2011). Die Maßnahmen sind zudem in den Unterrichtsalltag und die Schulkultur zu integrieren (vgl. Rosebrock & Nix, 2011, S. 131ff.). Laut Spinner (2004) sollten Lehrkräfte grundsätzlich bei jeder Textarbeit begleitend Lesetechniken vermitteln.

Als umfassendes Programm für die Sekundarstufe, das alle Ebenen der Leseförderung berücksichtigt, ist z. B. „Reading Apprenticeship“ (auch als „Reading for Understanding“ bekannt) von Schoenbach, Greenleaf, Cziko und Hurwitz (1999) zu nennen. Das Programm reicht von der Vermittlung der Erkenntnis, dass die eigene Lesekompetenz unzureichend ist, über das Training hierarchieniedriger und hierarchiehöherer Prozesse, die Reflexion über die Lektüre bis hin zum systematischen Wissensaufbau. Das ehrgeizig anmutende Programm wurde in den USA bereits mehrfach evaluiert und erwies sich als erstaunlich erfolgreich (Rosebrock & Nix, 2011; Schoenbach et al., 1999; Gaile, 2009). Gaile und Schoenbach (2006) brachten dieses Programm unter dem Namen „Lesen macht schlau“ nach Deutschland. Ein ebenfalls umfassendes Programm, das für den deutschen Sprachraum evaluiert wurde, ist „LekoLemo“ (Streblow, Holodynski & Schiefele, 2007; Streblow, Schiefele & Riedel, 2012). Es beinhaltet Strategievermittlung, Textverständnisübungen sowie Maßnahmen zur Motivationssteigerung und -aufrechterhaltung. Die Evaluation des aktuellen, überarbeiteten Programms aus dem Jahr 2012 ergab positive Effekte auf das Leseverständnis (nicht jedoch auf die Lesemotivation).

8.5 Fazit

In den vergangenen Abschnitten wurden verschiedene Ansatzpunkte und Maßnahmen zur Förderung von Leseverständnis und Lesekompetenz in der Sekundarstufe vorgestellt. Wichtig scheint bei der Auswahl der Maßnahmen vor allem, dass lernstandsabhängig unterschiedliche Maßnahmen sinnvoll sind. Bei Schülern, die z. B. aufgrund eines Migrationshintergrundes einen unzureichenden Wortschatz aufweisen, ist dieser entsprechend zu fördern, während bei Schülern, die Probleme mit basalen Le-

sekompetenzen haben, die phonologische Informationsverarbeitung sowie die Dekodierleistung und die Leseflüssigkeit trainiert werden sollten. Schüler, die die basalen Lesetechniken beherrschen, jedoch nicht motiviert sind zu lesen oder den Bedeutungsinhalt eines Textes nicht rekonstruieren und/oder das Gelesene nicht in einen sinnvollen Zusammenhang integrieren können, werden motivationsfördernde Maßnahmen, Strategietrainings sowie die Vermittlung von Textwissen und inhaltlichem Vorwissen empfohlen.

Für die Entscheidung, an welcher Stelle im Einzelfall anzusetzen ist, ist es notwendig herauszufinden, auf welchen Ebenen Defizite vorliegen. Hier wird erneut deutlich, dass ein standardisiertes Diagnoseinstrument aufzeigen könnte, ob Schwierigkeiten vor allem in Bezug auf basale Leseprozesse oder in Bezug auf das Textverständnis vorliegen.

Kapitel 9

Zusammenfassung und Implikationen aus Teil I

Im Folgenden werden der in den vorherigen Kapiteln beschriebene theoretische Teil der vorliegenden Arbeit und insbesondere die daraus gezogenen Implikationen für die Konstruktion von LESEN 6-7 und LESEN 8-9 zusammengefasst.

Eine Betrachtung des Lesekompetenzerwerbs zeigte zunächst, dass dieser sich nicht auf die Grundschulzeit beschränkt, sondern bereits davor beginnt und sich auch über die Sekundarschulzeit erstreckt. Mit zunehmender Automatisierung der basalen Lesekompetenzen des Rekodierens und Dekodierens stehen im Sekundarschulalter mehr kognitive Kapazitäten für inhaltliches Textverständnis zur Verfügung. Dadurch wird ein gezielter Einsatz von Lesestrategien möglich, wodurch wiederum die Informationsentnahme im Laufe der Sekundarschuljahre immer besser gesteuert und optimiert werden kann. Auch wenn die Leistungszuwächse in Bezug auf das Lesen in der Sekundarstufe nicht mehr so groß sind wie in der Grundschule, lassen sich dennoch Leistungssteigerungen beobachten.

Bei geübten Lesern findet schließlich eine aktive Text-Leser-Wechselwirkung statt, die sowohl Bottom-up- als auch Top-down-Prozesse impliziert. Hierarchieniedrigere, automatisierte Prozesse auf Wortebene, Satzebene sowie satzübergreifender Ebene führen zum Aufbau lokaler Kohärenz. Mithilfe hierarchiehöherer, bewusster und steuerbarer Prozesse kann darüber hinaus globale Kohärenz hergestellt werden. Wesentlich für die Kohärenzbildung sind Inferenzen, also über den expliziten Inhalt eines Textes hinausgehende Schlüsse unter Rückgriff auf bereits vorhandenes Wissen. Durch die genannten Prozesse entsteht eine mentale Textrepräsentation, wobei meist drei Repräsentationsformen unterschieden werden: (1) Die wörtliche Repräsentation oder Textoberfläche, (2) die propositionale Repräsentation oder Textbasis und (3) das über den expliziten Textinhalt hinausgehende Situationsmodell. Darüber hinaus begleiten gute Leser den Leseprozess zusätzlich auf der Metaebene. Diesbezüglich werden zwei weitere Verständnisebenen unterschieden, nämlich (4) das Erkennen genrespezifischer Textstrukturen sowie (5) das Erkennen rhetorischer Strategien des Autors und dessen Intention. Für einen Lesetest, der die Lesekompetenz von Sekundarschülern umfassend abbilden und dabei im gesamten Leistungsspektrum differenzieren soll, ist es daher erforderlich zusätzlich zur basalen Lesekompetenz auch das Textverständnis

einschließlich der verschiedenen Repräsentationsformen und der Metaebenen zu erfassen.

Trotz der Komplexität des Lesens und der vielfältigen daran beteiligten Prozesse sprechen die bisherigen Befunde dafür, dass es sich bei der Lesekompetenz um ein zweidimensionales Konstrukt handelt, wobei zwischen Dekodierfähigkeit und modalitätsunabhängigen Verstehensprozessen unterschieden werden kann. Bisweilen zeichnet sich darüber hinaus eine Geschwindigkeitskomponente als dritter Faktor ab, was jedoch noch genauer zu untersuchen ist. Einem Lesetest sollte daher auch kein weiter ausdifferenziertes Modell mit weiteren Skalen zugrunde gelegt werden. Empirische Befunde deuten zudem darauf hin, dass die auf den Abruf gelesener Worte aus dem mentalen Lexikon folgenden Verstehensprozesse beim Lesen weitgehend mit jenen beim Hörverstehen identisch sind. Entsprechend können schwache Leseleistungen aus schwacher Dekodierleistung, schwachem rezeptivem Verständnis oder einer Kombination aus beidem resultieren.

In Bezug auf schwache Leseleistungen bietet das DSM-5 die Möglichkeit, eine spezifische Lernstörung mit Defiziten im Bereich des Lesens zu diagnostizieren, während anhand des ICD-10 lediglich die mit der Rechtschreibleistung kombinierte Lese-Rechtschreibstörung als Diagnose gestellt werden kann. Bei der Diagnostik sollte unter anderem eine Beurteilung seitens der Schule herangezogen werden. Die diagnostische Kompetenz von Lehrkräften im Hinblick auf die Lesekompetenz erwies sich jedoch teilweise als problematisch. Sie fiel interindividuell sehr unterschiedlich aus und war generell in der Sekundarstufe niedriger ausgeprägt als in der Grundschule. In der Sekundarstufe zeigten sich sogar bei der Identifikation gravierender Leseverständnisdefizite Probleme. Für eine fundierte Diagnose sind daher auf jeden Fall auch standardisierte Lesetests notwendig. Diese können Lehrkräfte darüber hinaus generell auch im Schulalltag bei der Einschätzung der Lesekompetenz ihrer Schüler unterstützen. Standardisierte Lesetests sollten dabei stets auf einer Testtheorie basieren, aus der hervorgeht, welche Schlüsse die Itemantworten einer Person auf das interessierende Merkmal zulassen. Hierfür steht zum einen die auf Reliabilitätsaspekte fokussierende KTT zur Verfügung und zum anderen die IRT, deren Schwerpunkte Skalierbarkeit und Konstruktvalidität sind. Beide Theorien können sich komplementär ergänzen. Über die Reliabilität, Skalierbarkeit und Validität hinaus gibt es noch zahlreiche weitere Gütekriterien, die standardisierte Tests erfüllen sollten, z. B. Objektivität, Normierung und Ökonomie. Qualitativ hochwertige Lesetests können das Leseverständnis der Schüler objektiver und in Bezug auf eine größere Referenzgruppe beurteilen als dies den Lehrkräften durch Beobachtungen im Unterrichtsalltag möglich ist. Die einzelnen Teilprozesse des Lesens können in solchen Tests auf unterschiedliche Weise operationalisiert werden, wobei sich die Operationalisierung z. B. auf die Möglichkeit, den Test als Gruppentest einzusetzen, und somit auf die Testökonomie auswirkt. Für eine Integration von Lesetests in den Schulalltag sollten diese möglichst als Gruppentests einsetzbar sein. Auch für den Einsatz im Rahmen von großangeleg-

ten Forschungsprojekten ist die Einsetzbarkeit als Gruppentest ein großer Vorteil. Ein Überblick über derzeit verfügbare und gängige Lesetests zeigte, dass diesbezüglich für die Sekundarstufe bislang eine Lücke in den Diagnosemöglichkeiten bestand. Insbesondere als Gruppentest einsetzbare Tests, die über basale Leseprozesse hinaus auch höhere Verständnisebenen überprüfen, fehlten.

Genauer beschrieben wurde schließlich die Konzeption der PISA-Lesetests, da die mit diesen Tests erhobenen Befunde sowohl für die vorliegende Arbeit als auch generell für die Leseverständnisforschung von herausragender Bedeutung sind. Die Betrachtung der Befunde von PISA und weiterer Studien, die sich mit der Lesekompetenz befassten, zeigte u. a., auf welchem Niveau sich die Leseleistung der Zielgruppen der neu zu konstruierenden Lesetests befindet und welche Subgruppen der Zielgruppen sich hinsichtlich der Lesekompetenz unterscheiden lassen. Aus diesen Befunden geht zum einen hervor, welche Aspekte von Lesekompetenz die Tests erfassen sollten, um gut zwischen den Schülern der Zielgruppen differenzieren zu können, zum anderen können aus den Vorbefunden Hypothesen abgeleitet werden, anhand derer sich die Validität der Tests überprüfen lässt. Darüber hinaus sind die Befunde zu Differenzen zwischen verschiedenen Subgruppen bei der Ziehung der Normstichprobe und Repräsentativitätsprüfung der Normstichprobe zu berücksichtigen. Im Einzelnen zeigten sich Differenzen in der Lesekompetenz zwischen den Bundesländern, zwischen den Schularten, zwischen Schülern mit und ohne Migrationshintergrund sowie häufig auch zwischen den Geschlechtern (zugunsten der Mädchen). Insgesamt weisen die Ergebnisse auf Defizite deutscher Jugendlicher vor allem auf den höheren Verarbeitungsebenen, also z. B. beim Reflektieren, Bewerten und eigenständigen Interpretieren hin. Dies unterstreicht die Notwendigkeit, dass Lesetests, die das Leseverständnis in den mittleren Klassenstufen abbilden und Ansatzpunkte für Fördermaßnahmen aufzeigen sollen, auch die höheren Verständnisebenen einbeziehen sollten.

Die beschriebenen interindividuellen Differenzen in der Ausprägung der Lesekompetenz scheinen – wie ein Einblick in die Erforschung der Einflussfaktoren zeigte – multifaktoriell bedingt, wobei die Bedeutung der einzelnen Faktoren sowie zum Teil die Wirkrichtung nicht eindeutig geklärt sind, da vielfältige Interaktionen die Untersuchung erschweren. Insgesamt haben sowohl leserseitige Merkmale als auch das soziale Umfeld und Eigenschaften des Textes einen Einfluss darauf, wie gut eine Person einen Text verstehen kann. Abhängig davon, welche Defizite vorliegen und welche Ursachen dafür verantwortlich sind, sind entsprechende Fördermaßnahmen zu ergreifen. Bei Wortschatzdefiziten ist anders vorzugehen als bei Defiziten in Bezug auf die basalen Lesekompetenzen, und wieder anders sollten Schüler gefördert werden, die zwar über basale Lesekompetenzen verfügen, jedoch Verständnisprobleme oder motivationale Probleme beim Lesen haben. Daher sollte zunächst eine Diagnostik stattfinden, um vorliegende Defizite genau zu identifizieren. Lesetests für die Sekundarstufe, deren Ergebnisse eine solche Basis für gezielte Fördermaßnahmen bilden sollen, sollten demnach Hinweise darauf geben, wo beim Leser ein Defizit besteht.

Teil II

Konstruktion und empirische Erprobung von LESEN 6-7 und LESEN 8-9

Der nun folgende empirische Teil der vorliegenden Arbeit befasst sich mit der Konstruktion der beiden Lesetests LESEN 6-7 und LESEN 8-9 sowie mit deren empirischer Erprobung und Evaluation. Dabei wird zunächst die Zielsetzung der Testkonstruktion dargestellt und dann ein Überblick über das Projekt „LESEN – *L*esen *e*rmöglicht *Sinn*e*n*tnahme“ gegeben. Einige Vorbemerkungen zu den Datenerhebungen, zur statistischen Auswertung, zu den verwendeten Computerprogrammen sowie zu in den folgenden Kapiteln verwendeten Abkürzungen leiten sodann zur Beschreibung der Testkonstruktion über. Diese beginnt mit einer Erläuterung der theoretischen Fundierung der Tests und enthält weiter alle Schritte der Testkonstruktion von den ersten Testentwürfen über die Itemanalysen gemäß KTT und IRT sowie die Revisionen bis hin zu den Endversionen. Anschließend wird die umfangreiche Normierung dieser Endversionen beschrieben. Es folgen weitere KTT- und IRT-Analysen auf Basis der Normdaten und im Anschluss daran die Reliabilitätsanalysen und Validitätsanalysen für LESEN 6-7 und LESEN 8-9. Am Ende jedes Kapitels steht eine Diskussion der Ergebnisse und zuletzt erfolgt eine Zusammenfassung der gesamten empirischen Arbeit. Begonnen wird zunächst mit der Beschreibung der Zielsetzung.

Kapitel 10

Zielsetzung der Testkonstruktion

Im ersten Teil der vorliegenden Arbeit wurde deutlich, dass für die Sekundarstufe geeignete, standardisierte Tests fehlen, die das Leseverständnis umfassend abbilden und im gesamten Leistungsspektrum differenzieren können. Derartige Tests fehlten für den Einsatz im Schulalltag durch Lehrkräfte ebenso wie für die Forschung. Aus dieser Erkenntnis ergab sich für die Testkonstruktion die im Folgenden beschriebende Zielsetzung.

Als Zielgruppen wurden für LESEN 6-7 Schüler der Klassenstufen sechs und sieben und für LESEN 8-9 Schüler der Klassenstufen acht und neun definiert, wobei daraus ein Altersbereich von ca. 11 bis 14 Jahren bzw. ca. 13 bis 16 Jahren resultiert. Die Tests sollten für alle Schüler deutscher Regelschulen geeignet sein. In den genannten Klassenstufen fallen darunter Hauptschulen[6], Realschulen, Gymnasien und Gesamtschulen sowie in Bundesländern mit sechsstufiger Grundschule auch das letzte Grundschuljahr. Sonder- und Förderschulen sollten nicht berücksichtigt werden.

Die beiden Lesetests sollten analog aufgebaut sein und die Lesekompetenz erfassen. Wie im ersten Teil der Arbeit deutlich wurde, ist Lesekompetenz jedoch ein sehr umfassendes und komplexes Konstrukt, das unmöglich vollständig mit einem ökonomisch durchzuführenden und auswertbaren Test erfasst werden kann. Es wurde daher entschieden, auf die kognitiven Aspekte der Lesekompetenz, also das Leseverständnis, zu fokussieren, da dieses den Kern des Lesekompetenzkonstrukts bildet. Die Fähigkeit zum Leseverständnis kann jedoch nicht isoliert geprüft werden, denn die in einer Testsituation gezeigte Leseleistung ist stets von motivationalen und affektiven Faktoren beeinflusst. Dennoch sollte der Schwerpunkt von LESEN 6-7 und LESEN 8-9 auf den kognitiven Aspekten des Lesens liegen und das Leseverständnis möglichst umfassend abbilden. Bei der Testkonstruktion sollte entsprechend besonderer Wert auf eine gute theoretische Fundierung und die Berücksichtigung des aktuellen Forschungsstandes zum Leseverständnis gelegt werden. Die motivationalen und affektiven Aspekte fließen zwar zwangsläufig in die Testergebnisse mit ein, sind jedoch nicht expliziter Gegenstand der Tests und daher wird versucht, ihre Einflüsse so weit wie möglich zu minimieren.

[6] Zum Teil wurden die Hauptschulen in den letzten Jahren umstrukturiert und umbenannt, z. B. in Bayern in „Mittelschulen“. Zugunsten der Übersichtlichkeit wird jedoch in der vorliegenden Arbeit nur der Begriff „Hauptschule“ verwendet.

Um in allen vier Klassenstufen und allen genannten Schularten im gesamten Leistungsspektrum differenzieren zu können, wurde entschieden, neben der basalen Lesekompetenz auch das Leseverständnis auf höheren Ebenen zu erfassen. Aus den Testergebnissen sollte hervorgehen, wo genau Defizite im Leseverständnis eines Schülers liegen – beispielsweise schon bei den basalen Lesekompetenzen oder erst beim tiefergehenden Textverständnis – um dann gegebenenfalls angemessene Fördermaßnahmen ergreifen zu können.

Ein weiteres Ziel war es, dass LESEN 6-7 und LESEN 8-9 den gängigen Testgütekriterien genügen. Um dies bestmöglich erreichen zu können, sollten beide zur Verfügung stehenden Testtheorien (KTT und IRT) bei der Testkonstruktion berücksichtigt werden. Die Lesetests sollten nicht nur standardisiert, reliabel und valide, sondern auch normiert und ökonomisch handhabbar sein. Zudem sollten sie praktisch erprobt werden. Die Qualität der Tests, insbesondere die Reliabilität und die Validität, sollte nicht nur für Gruppenvergleiche, sondern auch für den Einsatz im Rahmen der Individualdiagnostik ausreichen. Um breit abgestützte und repräsentative Vergleichswerte zur Verfügung stellen zu können, wurde eine Normierung der Tests an einer großen, für Deutschland möglichst repräsentativen Stichprobe angestrebt.

Weiter sollten LESEN 6-7 und LESEN 8-9 als Gruppentests einsetzbar sein. Dies erleichtert eine Durchführung der Tests im Unterrichtsalltag und gibt Lehrkräften die Möglichkeit, auf ökonomische Weise einen Überblick über den Leistungsstand und interindividuelle Leistungsdifferenzen innerhalb ihrer Schulklasse zu erhalten. Beim Einsatz in der Forschung wird durch Gruppentests eine ökonomische Erfassung der Lesekompetenz bei großen Stichproben möglich. Um die Einsetzbarkeit im Schulalltag weiter zu erleichtern, wurde angestrebt, die Tests so zu gestalten, dass sie jeweils innerhalb einer Schulstunde (45 Minuten) durchführbar sind. Lesetests von längerer Dauer wären den schwächeren Schülern der Zielgruppe von LESEN 6-7 und LESEN 8-9 auch nicht zuzumuten.

Die Gestaltung von LESEN 6-7 und LESEN 8-9 und die Beschreibung der Tests in den Testmanualen sollten es einerseits Lehrkräften ermöglichen, die Tests selbst einzusetzen, auszuwerten und die Ergebnisse zu interpretieren sowie aus den Testergebnissen gegebenenfalls gezielte Fördermaßnahmen abzuleiten. Zugleich sollten Forscher andererseits alle Informationen erhalten, die sie benötigen, um die Tests zu Forschungszwecken einsetzen zu können.

Kapitel 11

Überblick über das Projekt

Über einen Zeitraum von zwei Jahren – von Juli 2009 bis Juni 2011 – fanden im Rahmen des Projektes „LESEN – *L*esen *e*rmöglicht *S*inn*en*tnahme“ die Testkonstruktion und die empirische Erprobung von LESEN 6-7 und LESEN 8-9 statt (s. Abb. 5). Nach einer Einarbeitung in den aktuellen Forschungsstand mittels einer Literaturanalyse der Gebiete Lesekompetenz und Testkonstruktion wurden im Dezember 2009 erste Testentwürfe erstellt. Diese wurden in mehreren Voruntersuchungen überprüft und mithilfe von Itemanalysen (basierend auf der KTT und teilweise zusätzlich der IRT) optimiert. Im März 2010 waren schließlich die beiden Endversionen fertiggestellt.

Die Normierung der Endversionen fand am Ende des Schuljahres 2009/2010 in mehreren deutschen Bundesländern statt. Anschließend wurde zu Beginn des Schuljahres 2010/2011 im süddeutschen Raum eine weitere Datenerhebung zur Validierung durchgeführt, im Rahmen derer LESEN 6-7 und LESEN 8-9 zusammen mit weiteren Lesetests und Fragebögen eingesetzt wurden. In der Mitte des Schuljahres 2010/2011 erfolgte außerdem an einer kleineren Stichprobe eine Wiederholungsmessung, um die Reliabilität auch anhand der Retestmethode bestimmen zu können.

Darauf folgte eine umfangreiche Auswertung der Normdaten, einschließlich einer kritischen Prüfung der Normstichprobe im Hinblick auf ihre Gesamtgröße sowie die Größe verschiedener Substichproben und ihre Repräsentativität. Zudem wurden Normtabellen erstellt und auf der Basis der Normdaten erneut Itemanalysen durchgeführt, um die in den Voruntersuchungen ermittelte Itemgüte an einer größeren Stichprobe zu replizieren. Hierzu wurden wiederum in erster Linie KTT-Kennwerte berechnet und teilweise darüber hinaus IRT-Analysen durchgeführt. Im Rahmen der Reliabilitätsanalysen wurden neben der Retestmethode auch verschiedene Konsistenzmethoden, sowie teilweise IRT-Methoden angewendet. Zur Validitätsprüfung wurden die Ergebnisse zusätzlicher Tests sowie die anhand zusätzlicher Fragebögen erhobenen Schulnoten und das Lehrerurteil bezüglich der Lesekompetenz herangezogen, um über die aus theoretischen Überlegungen abgeleitete Inhaltsvalidität hinaus auch die Konstrukt- und Kriteriumsvalidität bestimmen zu können. Weiter wurde geprüft, ob LESEN 6-7 und LESEN 8-9 in Bezug auf aus der Theorie abgeleitete Hypothesen zu erwartungskonformen Ergebnissen kommen und ob sie empirische Vorbefunde replizieren können.

Abschließend wurden die Tests auch im Hinblick auf weitere Gütekriterien betrachtet und das Manual erstellt. Im Sommer 2012 veröffentlichte der Hogrefe Verlag

LESEN 6-7 und LESEN 8-9 in der Reihe Hogrefe Schultests unter der Autorenschaft von Kerstin Bäuerlein, Wolfgang Lenhard und Wolfgang Schneider. Als Herausgeber fungierten Marcus Hasselhorn, Wolfgang Schneider und Ulrich Trautwein.

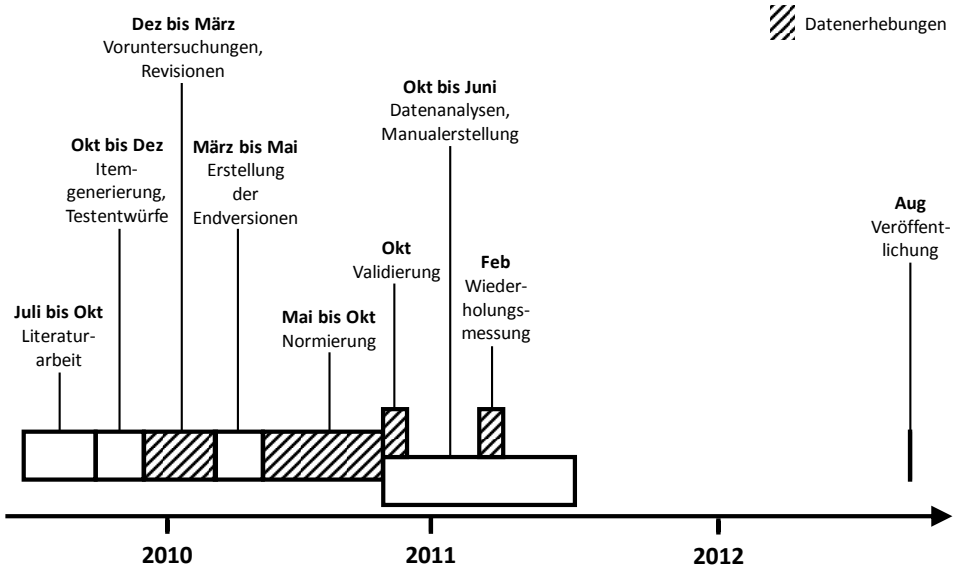


Abbildung 5. Zeitlicher Ablauf des Testkonstruktionsprojektes „LESEN – *L*esen *e*rmöglicht *S*inn*en*tnahme“.

Kapitel 12

Vorbemerkungen

Bevor die Beschreibung der Testkonstruktion beginnt, seien einige Anmerkungen zum generellen Vorgehen bei den Datenerhebungen, zu den eingesetzten Testmaterialien, zu den statistischen Analysen sowie zu den verwendeten Computerprogrammen vorangestellt. Zudem werden einige im weiteren Verlauf der Arbeit verwendete Abkürzungen eingeführt.

Datenerhebungen. Vor allen Datenerhebungen wurde die Genehmigung der jeweils zuständigen Behörde (Bezirksregierung bzw. Kultusministerium) eingeholt. Anschließend wurden Schulleitungen kontaktiert, über das Projekt informiert und um Teilnahme gebeten. Bei Teilnahmebereitschaft erhielten sie schriftliche Lehrer-, Eltern- und Schülerinformationen über Ziel und Ablauf der Untersuchung sowie über die Freiwilligkeit der Teilnahme. Zudem wurde den Lehrkräften eine schriftliche Rückmeldung über das Abschneiden der eigenen Schüler im Vergleich zu allen anderen teilnehmenden Schülern der entsprechenden Klassenstufe und Schulart angeboten. Die Rückmeldung erfolgte anonymisiert.

Bei Vorliegen einer schriftlichen Einverständniserklärung der Eltern (und bei Schülern über 14 Jahren auch von diesen selbst) wurde ein Testtermin vereinbart. Die Untersuchungen wurden als Gruppentests vormittags während der regulären Unterrichtszeit im normalen Klassenzimmer bzw. bei einer Voruntersuchung, die computergestützt stattfand, im PC-Raum einer Schule durchgeführt. Die Durchführung erfolgte standardisiert, indem von einem mit dem Test gut vertrauten Testleiter ein vorgegebener Instruktionstext wörtlich wiedergegeben wurde. Die Schüler erhielten nach der Testung ein kleines Dankeschön (z. B. einen Stift oder einen Radiergummi).

Testmaterialien. Die Endversionen von LESEN 6-7 und LESEN 8-9 sowie die bei der Validierung eingesetzten Tests LGVT 6-12 und WLST 7-12 sind beim Hogrefe Verlag veröffentlicht. Informationen zu dem bei den sechsten Klassenstufen ebenfalls im Rahmen der Validierung herangezogenen Metakognitionstest finden sich bei Lingel, Neuenhaus, Artelt und Schneider (2010). Weitere verwendete Items werden in den entsprechenden Abschnitten beschrieben.

Statistische Auswertung. Falls nicht gesondert vermerkt, gilt für alle Berechnungen ein Signifikanzniveau von $\alpha = .05$. Dieser Wert ist per Konvention festgelegt (vgl. Bortz & Schuster, 2010, S. 101).

Computerprogramme. Für die Erstellung der computerbasierten Pilotversion eines Subtests wurden *Java SE Runtime Environment* (JRE) Version 6 und *JavaFX Runtime* Version 1 verwendet. Die Berechnungen im Rahmen der KTT-Itemanalysen, die Gruppenvergleiche im Rahmen der Validitätsanalysen sowie einige Berechnungen im Rahmen der Faktorenanalysen erfolgten mit der Statistiksoftware *SPSS* (Versionen 15.0.1, 18 und 20x32). Weitere faktorenanalytische Berechnungen wurden mit *MPlus5* durchgeführt. Für die Berechnungen im Rahmen der IRT-Analysen wurden darüber hinaus *ConQuest* (Version 1.0.0.1) und *R* (Version i386 2.15.1) verwendet. Zusätzlich wurde für verschiedene Berechnungen *Microsoft Office Excel* (Version 2010) herangezogen. Die Berechnung der Effektstärken erfolgte mit *G*Power* (3.1; Faul, Erdfelder, Lang & Buchner, 2007; Faul, Erdfelder, Buchner & Lang, 2009).

Abkürzungen. Im Folgenden werden die verschiedenen Schularten in den Tabellen abgekürzt. „HS“ steht für Hauptschule, „RS“ für Realschule, „GYM“ für Gymnasium und „AN“ für Andere (Werkrealschulen, M-Zug-Klassen[7]). Ist von der Gesamtstichprobe oder „Gesamt“ die Rede, sind – sofern nicht anders vermerkt – alle diese vier Gruppen eingeschlossen. Das Gesamtergebnis über beide Subtests von LESEN 6-7 bzw. LESEN 8-9 wird dagegen in Tabellen mit „GES“ abgekürzt.

[7] Werkrealschulen und M-Zug-Klassen sind Haupt- bzw. Mittelschulzweige in Baden-Württemberg bzw. Bayern, die den Abschluss der Mittleren Reife ermöglichen.

Kapitel 13

Testkonstruktion

Im Folgenden wird das Vorgehen bei der Testkonstruktion von der theoretischen Fundierung über die Testentwürfe und Revisionen bis zu den Endversionen im Detail dargestellt.

13.1 Theoretische Fundierung

Aus den im ersten Teil der Arbeit dargestellten Theorien und Befunden ergaben sich zahlreiche Implikationen für die theoretische Fundierung von LESEN 6-7 und LESEN 8-9. Diese werden in den folgenden Abschnitten noch einmal zusammengefasst.

Aus einer Vielzahl von Studien geht hervor, dass auch in der Sekundarstufe bei vielen Schülern im Hinblick auf basale Lesekompetenzen Defizite vorliegen (vgl. Artelt et al., 2001; Klicpera, Schabmann & Gasteiger Klicpera, 1993; Stanat & Schneider, 2004). Daher sollten die neuen Tests auch die basale Lesekompetenz berücksichtigen. Darüber hinaus sollten sie, um das Leseverständnis umfassend abzubilden und auch im oberen Leistungsbereich differenzieren zu können, die Textverarbeitung auf hierarchiehöheren Ebenen überprüfen. Trotz der Annahme unterschiedlicher Ebenen der Verarbeitungstiefe werden jedoch innerhalb der Verständniskomponente keine distinkten Dimensionen angenommen (vgl. Kap. 4.3).

Da verschiedene Studien zeigten, dass sowohl isolierte Schwächen in der basalen Lesekompetenz als auch isolierte Schwächen der Verständnisleistung auftreten können (Catts et al., 2006; Stothard & Hulme, 1995; Ennemoser et al., 2012), sollte bei der Konstruktion von LESEN 6-7 und LESEN 8-9 darauf geachtet werden, diese beiden Aspekte möglichst separat zu erfassen, um im Einzelfall feststellen zu können, wo genau Probleme bestehen und somit Ansatzpunkte für gezielte, individuelle Förderung aufzuzeigen. Dass keine vollständige Separierung möglich ist, wurde bereits in Kapitel 5.3.3 deutlich. Dennoch sollte eine vollständige Konfundierung (wie z. B. bei Lückentexten) vermieden werden.

Aus diesen Befunden und Überlegungen ergab sich somit, dass jeder der zwei zu konstruierenden Lesetests zwei Subtests enthalten sollte: einen zur Erfassung der basalen Lesekompetenz und einen zur Erfassung des Textverständnisses. Für die Erfassung der basalen Lesekompetenz wurde eine Satzleseaufgabe mit knapper Zeitbegrenzung gewählt, denn Satzleseaufgaben können im Gegensatz zu Wortleseaufgaben

auch in der Sekundarstufe noch zwischen guten und schlechten Lesern differenzieren. Darüber hinaus bewährten sie sich bereits bei anderen standardisierten Tests für die Erfassung der basalen Lesekompetenz in der Sekundarstufe (z. B. beim SLS 5-8 Auer et al., 2005, s. Kap. 5.3.4).

Auf Lautleseaufgaben, mit deren Hilfe z. B. die sinnangemessene Intonation im Sinne der Leseflüssigkeit in der Definition von Nix (2011, S. 61; s. Kap. 2) erfasst werden könnte, musste bei den neuen Tests verzichtet werden. Da ein Gruppentest konstruiert werden sollte. Stille Leseaufgaben haben jedoch den Vorteil, dass diese ökologisch valider sind, da stilles Lesen im Alltag den Regelfall darstellt (W. Lenhard, 2013, S. 88). Zudem ist in der Sekundarstufe vor allem die Lesegeschwindigkeit – im Gegensatz zur Lesegenauigkeit – für Leistungsunterschiede ausschlaggebend und diese kann auch über leises Lesen geprüft werden (vgl. W. Lenhard, 2013, S. 86).

Um auch im oberen Leistungsbereich differenzieren zu können, sollte das tiefergehende Textverständnis einschließlich globaler Kohärenz und vorwissensbasierter Inferenzbildung überprüft werden. Daher war es nötig, längere Texte vorzulegen, die Kohärenzbildung über größere Textabschnitte hinweg erfordern, und hierzu Verständnisfragen zu stellen, die sich auf die verschiedenen beschriebenen Ebenen des Textverständnisses beziehen (vgl. Irwin, 1991, S. 195). Bei den Fragen soll unterschieden werden zwischen Fragen, deren Beantwortung Wortverständnis im Kontext erfordert, Fragen, deren Beantwortung lokale Kohärenzbildung erfordert, und Fragen, deren Beantwortung globale Kohärenzbildung erfordert. Weiter soll unterschieden werden zwischen Fragen, für deren Beantwortung lediglich eine textbasierte (oberflächliche oder propositionale) Repräsentation aufgebaut werden muss, und Fragen, für deren Beantwortung ein Situationsmodell generiert werden muss. Darüber hinaus soll das Textverständnis auf den Metaebenen geprüft werden, indem Fragen zu genrespezifischem Textwissen – einschließlich Superstrukturen – sowie zu Darstellungsstrategien – einschließlich rhetorischer Strategien und Autorenintention – gestellt werden.

Da die Anforderungen an den Leser sowie die Erwartungen des Lesers an den Text abhängig vom Textgenre unterschiedlich sein können (Graesser et al., 1997), erschien es für die umfassende Überprüfung des Leseverständnisses zudem sinnvoll, verschiedene Textgenres einzubeziehen (vgl. auch FLVT 5-6; Souvignier, Hasselhorn, Schneider & Marx, 2009). Es wurden daher narrative und expositorische Texte ausgewählt, da dies die zwei Genres sind, welche im Allgemeinen in der Literatur grob unterschieden werden (vgl. Rosebrock & Nix, 2011, S. 76). Daraus ergab sich zudem ein Mittelweg zwischen rein literarischem und rein informationsfokussiertem Leseverständnis. Die Darbietung längerer Texte dieser Genres entspricht dabei bei der anvisierten Zielgruppe auch authentischen Lesesituationen im schulischen und außerschulischen Kontext. Somit ergibt sich ein Subtest zur Erfassung des Textverständnisses, der aus zwei Texten verschiedener Genres besteht. Zu jedem dieser Texte sollten alle genannten Verständnisebenen abgefragt werden.

Aufgrund der Eindimensionalitätsannahme für das Leseverständnis wird davon ausgegangen, dass dem Verständnis verschiedener Textgenres eine gemeinsame Verständnisdimension zugrunde liegt, und dass die Verständnisleistung bezüglich beider genreverschiedener Texte zu einem Gesamtergebnis für einen Subtest zur Erfassung des Textverständnisses zusammengefasst werden können sollte. Bei einer sehr großzügigen Zeitbegrenzung sollte die basale Lesekompetenz das Ergebnis hier nicht stark beeinflussen, da in der Sekundarschule diesbezüglich hauptsächlich die Dekodiergeschwindigkeit eine differenzierende Rolle spielt.

Wie bereits erläutert, kann das Leseverständnis nicht gänzlich isoliert von motivationalen und affektiven Aspekten erfasst werden, dennoch sollen letztere nicht explizit Gegenstand der Tests sein. Die Fokussierung von LESEN 6-7 und LESEN 8-9 auf die kognitiven, verständnisbezogenen Aspekte der Lesekompetenz wird auch in Abbildung 6 deutlich, welche die Komponenten des Leseverständnisses zusammenfasst, die aufgrund der beschriebenen theoretischen Überlegungen bei der Konstruktion der neuen Lesetests berücksichtigt werden sollten.

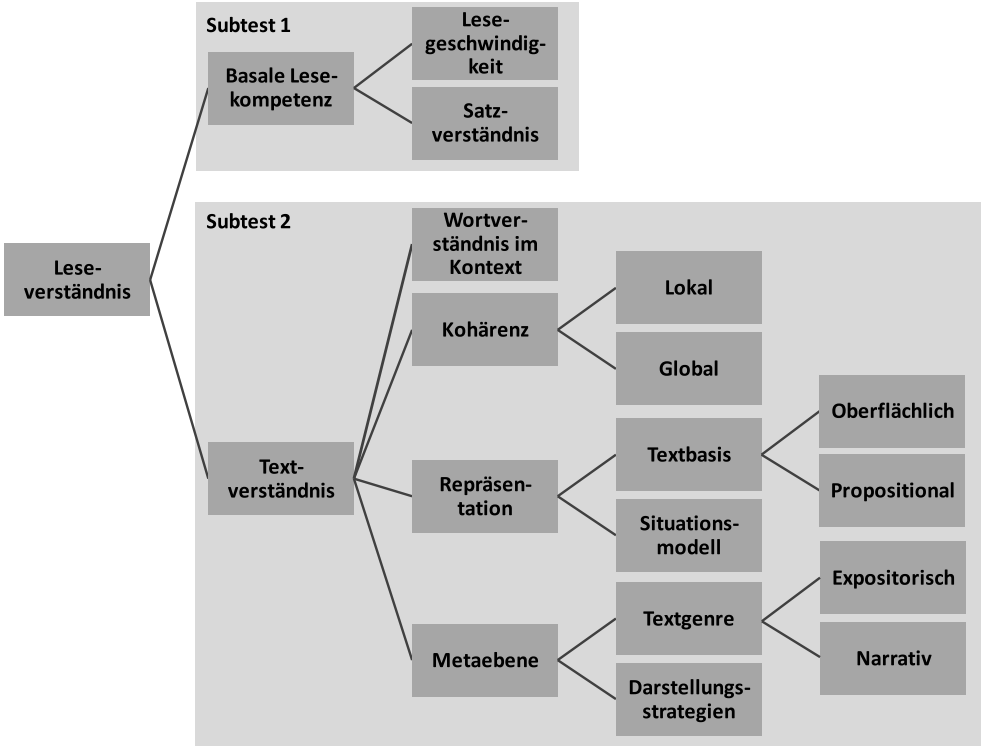


Abbildung 6. Aufbau der Lesetests LESEN 6-7 und LESEN 8-9.

13.2 Testentwürfe

Die im vergangenen Kapitel beschriebenen Implikationen aus den Theorien und Befunden zum Leseverständnis bildeten die Basis für erste Testentwürfe. Es wurden zwei Aufgabentypen erstellt: Eine erste Aufgabe in Form einer Satzleseaufgabe zur Erfassung der basalen Lesekompetenz und eine zweite Aufgabe mit Texten und dazugehörigen Verständnisfragen zur Erfassung des Textverständnisses.

Subtest 1: Basale Lesekompetenz (BLK). Für die Satzleseaufgabe zur Erfassung der basalen Lesekompetenz wurde zunächst eine Liste einfacher, kurzer Sätze (z. B. „Alle Häuser sind rund“) erstellt. Die Schüler sollten diese Sätze so schnell wie möglich lesen und entscheiden, ob sie inhaltlich richtig oder falsch sind. Es handelt sich also um ein gebundenes Antwortformat mit zwei Antwortalternativen. Die Sätze sollten so einfach sein, dass sie ohne Zeitdruck von allen Schülern der Zielgruppe korrekt beurteilt werden können. Der Schwerpunkt der Aufgabe sollte auf der Geschwindigkeitskomponente liegen. Die inhaltliche Beurteilung sollte lediglich sicherstellen, dass die Sätze auch tatsächlich gelesen werden. Es wurden für einen ersten Testentwurf 140 Sätze zusammengestellt, von denen 70 inhaltlich richtig und 70 inhaltlich falsch waren. Es wurde vorgesehen, denselben Subtest BLK für beide Tests (LESEN 6-7 und LESEN 8-9) zu verwenden, daher wurde nur ein Testentwurf erstellt.

Subtest 2: Textverständnis (TV). Für den Subtest TV wurden zwei expositorische und zwei narrative Texte ausgewählt, die jeweils 738 bis 895 Wörter umfassten. Bei der Auswahl der Texte wurde darauf geachtet, dass die in den Texten behandelten Themen „geschlechtsneutral“ und der Zielgruppe im Allgemeinen nicht sehr vertraut sind, damit nicht einzelne Schüler oder Subgruppen von Schülern bevorteilt werden. Die expositorischen Texte „Tiefsee“ und „Koboldmakis“ wurden neu verfasst. Bei den narrativen Texten wurde auf bereits existierende Texte zurückgegriffen. Es wurde jedoch bei der Auswahl darauf geachtet, dass das Urheberrecht schon erloschen war, um nötigenfalls Änderungen an den Texten vornehmen zu können. Die Wahl fiel auf „Der geheilte Patient“ von Johann Peter Habel und „Gedächtniskunst“ von Arkadij Awertschenko.

Zu jedem der Texte wurden 40 Verständnisfragen formuliert. Durch diese relativ große Anzahl von Verständnisfragen sollte die Itemselektion erleichtert werden, indem eine große Auswahl zur Verfügung gestellt wurde. Die Fragen bezogen sich möglichst gleichmäßig verteilt auf die verschiedenen Verständnisebenen (s. Abb. 6). Es wurde auch für diesen Subtest ein gebundenes Antwortformat gewählt. Durch die Vorgabe von vier Distraktoren (also einer relativ großen Anzahl) sollte die Ratewahrscheinlichkeit gering gehalten werden. Weiter wurde das SC-Format gewählt, bei dem explizit jeweils nur eine der fünf Antwortalternativen korrekt ist. Dieses Format dürfte den Schülern am vertrautesten sein und sollte somit Fehler aufgrund einer missverstandenen Instruktion vermeiden.

Zur Absicherung des Verständnisses der Instruktionen, wurde dem Subtest TV ein kurzer Text mit zwei Beispielitems vorangestellt (vgl. Empfehlung von J. Rost, 2004, S. 76). Ein Beispielitem bestand dabei aus einer Frage, deren Antwort wörtlich im Text zu finden war und das andere Beispielitem bestand aus einer Frage, für deren Beantwortung Inferenzbildung notwendig war. Damit sollte den Schülern verdeutlicht werden, dass die Antworten zum Teil wörtlich im Text zu finden sind und zum Teil Schlussfolgerungen und das Heranziehen von Vorwissen für die korrekte Beantwortung erforderlich sind.

Weiter wurde die Variante gewählt, dass die Schüler die Möglichkeit haben, während der Bearbeitung der Fragen zum Text zurückzublättern und noch einmal nachzulesen. Diese Variante wird gegenüber der Variante, dass die Möglichkeit des Zurückblätterns nicht besteht, bevorzugt, da sie für ökologisch valider gehalten wird (vgl. Weitsenfelder & Hofer, 2012). Es kommt in der Realität nämlich selten vor, dass die Möglichkeit, etwas noch einmal nachzulesen, nicht gegeben ist. Zudem soll der Test das Leseverständnis bzw. die Lesekompetenz und nicht etwa zusätzlich die Gedächtnisleistung prüfen.

Die Itemselektion sowie die Zuordnung der Texte zu den Klassenstufen bzw. zu den beiden Tests sollte erst auf der Basis von Erkenntnissen aus Voruntersuchungen erfolgen, welche im Folgenden beschrieben werden.

13.3 Voruntersuchungen und Revisionen

Alle Voruntersuchungen fanden im Zeitraum von Dezember 2009 bis März 2010 statt. Aus praktisch-organisatorischen und ökonomischen Gründen beschränkten sie sich auf den Regierungsbezirk Unterfranken. Die Subtests wurden in den Voruntersuchungen separat überprüft. Für den Subtest BLK wurde eine Voruntersuchung durchgeführt, für den Subtest TV fanden drei Voruntersuchungen statt.

13.3.1 Voruntersuchung zum Subtest BLK

Der Subtest BLK sollte für LESEN 6-7 und LESEN 8-9 identisch sein, weshalb die Voruntersuchung für beide Tests gemeinsam stattfand. Im Folgenden wird zunächst das methodische Vorgehen dargestellt, und anschließend werden die Ergebnisse berichtet.

13.3.1.1 Methode

Bevor auf Itemanalyse und Itemselektion eingegangen wird, werden zunächst die Art der Datenerhebung und die Stichprobe beschrieben.

Datenerhebung. Die Voruntersuchung für den Subtest BLK wurde computergestützt durchgeführt, um neben der Korrektheit der Antworten auch die Antwortzeiten auf-

zeichnen zu können. Die Schüler arbeiteten jeweils alleine an einem PC. Die Sätze wurden einzeln und in zufälliger Reihenfolge auf dem Bildschirm eingeblendet. Zusätzlich erschienen darunter die zwei Schaltflächen „richtig“ und „falsch“ (s. Abb. 7). Die Schüler sollten jeweils so schnell wie möglich per Mausklick auf eine der beiden Schaltflächen entscheiden, ob der dargebotene Satz inhaltlich richtig oder falsch ist. Nach dem Klicken folgte der nächste Satz.

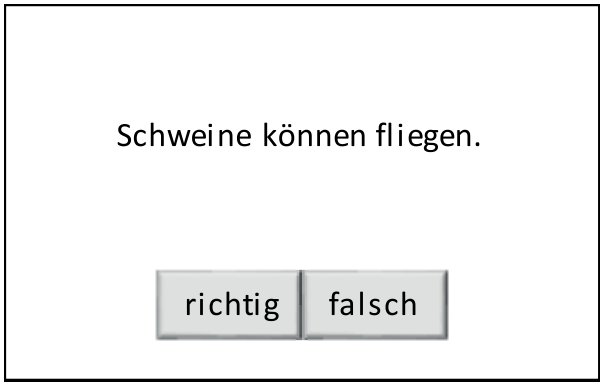


Abbildung 7. Exemplarische Bildschirmansicht bei der Voruntersuchung zum Subtest BLK.

Stichprobe. Die Stichprobe umfasste 41 Schüler. Da es sich als schwierig erwies, Schulen zu finden, die nicht nur bereit waren, die Untersuchung zu unterstützen, sondern zugleich einen dafür ausreichend gut ausgestatteten PC-Raum zur Verfügung hatten, beschränkte sich die Voruntersuchung auf lediglich eine Realschulklasse der sechsten Klassenstufe ($n = 16$) sowie eine Realschulklasse der neunten Klassenstufe ($n = 25$).

Itemanalyse und Itemselektion. Beim Subtest BLK fand die Itemselektion anhand der Lösungszeit und Korrektheit der Reaktionen der Schüler auf die einzelnen Items statt. Zwar gibt es auch verschiedene Ansätze, IRT-Methoden für Speedtests zu verwenden (vgl. z. B. Jansen, 1997a, 1997b, 2003; Jansen & Glas, 2005; Jansen, 2007), jedoch ist es im vorliegenden Fall nicht nötig, ein Testmodell zugrunde zu legen, da in der Endversion die Anzahl der korrekt bearbeiteten Items innerhalb einer vorgegebenen Zeit bereits eine metrische Variable darstellen wird (vgl. J. Rost, 2004, S. 44).

So wurden zunächst deskriptive Statistiken für die Fehlerzahlen und Antwortzeiten berechnet. Auf Basis der Überlegung, dass sich Leistung (Performanz) als geleistete Arbeit pro Zeiteinheit ausdrücken lässt, wurden weiter für alle Versuchspersonen über folgende Formel Leseperformanzquotienten (LPQ) für die Items ermittelt:

$$LPQ = \frac{\text{Korrektheit der Lösung } (1 = \text{richtig}, 0 = \text{falsch})}{\text{Lösungszeit (in Sekunden)}}$$

Im nächsten Schritt erfolgte die Itemselektion. Dabei wurden zunächst alle Items, die von fünf oder mehr Schülern falsch beantwortet worden waren, eliminiert und aus den verbleibenden Items die 100 Items mit den höchsten LPQ-Werten ausgewählt. Items mit hohem LPQ-Wert wurden am schnellsten korrekt beantwortet. Über die Analyse der Lösungszeiten für die ausgewählten 100 Items konnte zudem bestimmt werden, wie viel Zeit die schnellsten Schüler für die Bearbeitung der ausgewählten 100 Items benötigt hatten. Daraus wurde eine maximale Bearbeitungsdauer abgeleitet, denn es sollte möglichst wenigen oder gar keinen Schülern gelingen, alle Sätze in der vorgegebenen Zeit zu bearbeiten.

13.3.1.2 Ergebnisse

Im Folgenden wird zunächst auf die deskriptive Statistik eingegangen und anschließend werden die Itemselektion und die Bestimmung der Zeitbegrenzung für die Endversion des Subtests BLK dargestellt.

Deskriptive Statistik und Itemselektion. Die deskriptive Statistik sowie die LPQ-Werte für alle 140 Items befinden sich in Anhang A in Tabelle 28. Im Mittel wurden die Items 1.3 mal falsch beantwortet. Auslassungen waren nicht möglich, da das nächste Item erst dargeboten wurde, wenn für das vorherige „richtig“ oder „falsch“ angeklickt worden war. Die mittlere Bearbeitungszeit betrug 4.07 Sekunden mit einer Standardabweichung von 1.64 Sekunden. Die minimale Bearbeitungszeit betrug über alle Items gemittelt 1.69 Sekunden, die maximale Bearbeitungszeit 9.5 Sekunden. Der über die Gesamtstichprobe und alle Items gemittelte LPQ-Wert lag bei 0.32 mit einer Standardabweichung von 0.15, einem Minimum von 0.05 und einem Maximum von 0.80.

Im Rahmen der Itemselektion wurden acht Items eliminiert, da sie von fünf oder mehr Schülern nicht korrekt gelöst worden waren. Aus den übrigen Items wurden die 100 Items mit den höchsten LPQ-Werten ausgewählt. Diese 100 Items bilden nach LPQ-Wert absteigend angeordnet (also leichtestes Item an erster Stelle, schwerstes Item an letzter Stelle) die Endversion des Subtests BLK.

Maximale Bearbeitungsdauer. Als maximale Bearbeitungsdauer für die Endversion des Subtests BLK wurden drei Minuten festgelegt. Diese Zeitgrenze liegt etwas unterhalb der Zeit, die die schnellsten Schüler bei der Voruntersuchung für die ausgewählten 100 Items benötigten.

13.3.2 Erste Voruntersuchung zum Subtest TV

Die erste Voruntersuchung zum Subtest TV fand für alle vier im Rahmen der Erstellung der Testentwürfe ausgewählten Texte separat statt. Auch hier wird das methodische Vorgehen beschrieben, bevor auf die Ergebnisse eingegangen wird.

13.3.2.1 Methode

Im Folgenden werden zunächst die Datenerhebung und die Stichprobe betrachtet und anschließend wird auf die Datenanalysen eingegangen. Letztere umfassen deskriptive Analysen und Verteilungsanalysen sowie Itemanalysen und Reliabilitätsanalysen. Anschließend werden die Experten- und Zielgruppenbefragung sowie die Itemrevision dargestellt.

Datenerhebung. Die Datenerhebungen fanden hauptsächlich in Realschulklassen statt, um zu sehen, wie gut die Items etwa im mittleren Leistungsbereich der Zielgruppen funktionieren. Alle Texte und Verständnisfragen wurden dort in den Klassenstufen sechs und neun vorgelegt. Um zusätzlich einen Eindruck zu bekommen, wie die Items in den Randbereichen des Leistungsspektrums der Zielgruppen differenzieren, wurde der Text, der sich nach den ersten Datenerhebungen als der schwerste erwiesen hatte („Tiefsee“), zusätzlich in einer sechsten Hauptschulklasse durchgeführt und der Text, der sich nach den ersten Datenerhebungen als der leichteste erwiesen hatte („Der geheilte Patient“), wurde zusätzlich in einer neunten Gymnasialklasse durchgeführt.

Bei der Datenerhebung erhielt jeder Schüler einen Text mit 40 Verständnisfragen. Eine feste Zeitbegrenzung wurde für die Bearbeitung nicht festgelegt, denn alle Schüler sollten die Möglichkeit haben, den Text vollständig zu lesen und alle Items zu bearbeiten. Als grobe Richtlinie wurde eine Bearbeitungsdauer von einer Schulstunde vorgegeben.

Stichprobe. Die Stichprobe umfasste insgesamt 343 Schüler. Davon waren 138 Realschüler der sechsten Klassenstufe, 173 waren Realschüler der neunten Klassenstufe, 18 waren Hauptschüler der sechsten Klassenstufe und 14 waren Gymnasiasten der neunten Klassenstufe. Die Verteilung der Schüler auf die Texte kann Tabelle 3 im Ergebnisteil entnommen werden.

Datenanalyse und Itemselektion. Zunächst wurden deskriptive Statistiken und die Rohwertverteilungen für die einzelnen Klassenstufen und Schularten betrachtet, um einen ersten Eindruck von der Gesamtschwierigkeit der Items jedes Textes zu erhalten. Anschließend wurden Itemanalysen durchgeführt. Diese schlossen neben Distraktorenanalysen auch Schwierigkeits- und Trennschärfeanalysen ein sowie die Bestimmung des Selektionskennwertes für jedes Item. Bei der Itemselektion wurde über die statistischen Kennwerte hinaus stets die inhaltliche Bedeutung der einzelnen Items für

die Abbildung des Konstrukts berücksichtigt, um nicht bessere Kennwerte auf Kosten inhaltlicher Validität zu erzielen.

Ziel der Distraktorenanalysen war es, zu sehen, ob alle Distraktoren gleichmäßig angekreuzt wurden oder ob manche selten oder möglicherweise gar nicht angekreuzt wurden. Zudem kann die Betrachtung der Distraktoren helfen, herauszufinden, ob eine Antwortoption missverständlich formuliert, zu leicht oder auch zu schwer sein könnte, und ob eine Überarbeitung zur Verbesserung einzelner Items führen könnte (vgl. Lienert & Raatz, 1998, S. 123f.). So kann verhindert werden, dass Testitems von hoher inhaltlicher Validität nur aufgrund ungünstiger Antwort- oder Distraktorenformulierungen aus dem Test eliminiert werden. Entsprechend wurde für jedes Item eine Häufigkeitsverteilung der Antwortalternativen als Histogramm erstellt und betrachtet, welche Distraktoren nur sehr selten oder gar nicht angekreuzt wurden und somit eventuell einer Überarbeitung bedürfen. Da Distraktorenanalysen jedoch nur Hinweise auf möglicherweise schlechte Formulierungen geben können, wurden stets die Itemschwierigkeiten und -trennschärfen mitberücksichtigt.

Zur Bestimmung der Itemschwierigkeit wurde für jedes Item der Schwierigkeitsindex p ermittelt. Dieser entspricht bei dichotomen Items dem Itemmittelwert und nimmt einen Wert zwischen 0 und 1 an. Ein Wert von $p = 1$ bedeutet, dass alle Schüler das Item korrekt lösten, und ein Wert von $p = 0$ bedeutet, dass kein Schüler das Item korrekt löste. Itemschwierigkeiten zwischen $p = .20$ und $p = .80$ gelten als mittelschwer (Fisseni, 2004, S. 80) und werden in der Regel bevorzugt (Bortz & Döring, 2006, S. 219). Zu leichte oder zu schwere Items werden eliminiert. Mittlere Itemschwierigkeiten (um $p = .50$) werden auch deshalb angestrebt, weil eine mittelhohe Schwierigkeit die Wahrscheinlichkeit für eine hohe Trennschärfe erhöht (vgl. Fisseni, 2004, S. 35f.). Da im vorliegenden Fall jedoch insgesamt ein möglichst großes Spektrum an Schwierigkeiten vorhanden sein soll, um in allen Leistungsbereichen differenzieren zu können, wurden zunächst lediglich Items mit extremen Schwierigkeitswerten von $p > .90$ und $p < .10$ eliminiert.

Die Trennschärfe stellt die Korrelation eines Items mit dem Gesamtergebnis eines Tests dar (Bortz & Döring, 2006, S. 219). Da es sich hier um dichotome Items handelt, wurden punktbiserale Korrelationen durchgeführt. Die Korrelationen wurden „part-whole korrigiert“ (vgl. Moosbrugger, 2008a, S. 82). In der Regel gelten Werte von $r_{it} < .30$ als niedrig, Werte von $r_{it} = .30$ bis $r_{it} = .50$ als mittelhoch und Werte von $r_{it} > .50$ als hoch (z. B. Fisseni, 2004, S. 80). Meist wird für jedes Item eine Mindesthöhe der Trennschärfe von $r_{it} = .30$ gefordert (z. B. Fisseni, 2004; Lienert & Raatz, 1998). Da es sich an dieser Stelle um eine erste Voruntersuchung handelt, wurde das Selektionskriterium zunächst niedriger angesetzt, um die Möglichkeit offen zu halten, bei inhaltlich wichtigen Items zunächst durch Umformulierungen oder neue Distraktoren in der nächsten Voruntersuchung eine Verbesserung zu erzielen. So wurden nur Items mit einer sehr niedrigen Trennschärfe ($r_{it} < .20$) eliminiert. Zusätzlich wurde bei der Selektion der Items die Signifikanz der Trennschärfen als Kriterium

berücksichtigt. Items mit nicht signifikant von Null abweichender Trennschärfe wurden ebenfalls eliminiert. Da die Items im gesamten Fähigkeitsspektrum der Probanden differenzieren und weder sehr schwere noch sehr leichte Items aufgrund geringerer Streuung und damit einhergehend geringerer Trennschärfe sofort ausgeschlossen werden sollten, wurde zusätzlich zur Trennschärfe der Selektionskennwert (SK) berücksichtigt. Der SK ergibt sich aus der Trennschärfe eines Items dividiert durch die doppelte Standardabweichung und führt dazu, dass die Trennschärfe von Items mit extrem hohen oder extrem niedrigen Schwierigkeitsindizes nach oben korrigiert wird (Lienert & Raatz, 1998, S. 117f.).

Zur Prüfung der Reliabilität wurde schließlich die interne Konsistenz betrachtet. Da es sich im vorliegenden Fall um dichotome Items handelt, wurde hierfür die KR-20-Formel verwendet (vgl. Lienert & Raatz, 1998, S. 193).

Zielgruppen- und Expertenbefragung. Über die statistische Datenanalyse hinaus erfolgte eine Zielgruppen- und Expertenbefragung, um für die Itemüberarbeitung Hinweise und Änderungsvorschläge hinsichtlich der Texte und der Verständnisfragen zu erhalten. So wurden die Schüler nach Testende zur Schwierigkeit des Textes und der Items befragt und gebeten, es mitzuteilen, falls ihnen Fehler oder missverständliche Formulierungen aufgefallen sind oder sie sonstige Verbesserungsvorschläge haben. Die Deutschlehrkräfte wurden gebeten, während der Datenerhebungen in den Klassen den Test genau zu betrachten und gegebenenfalls Änderungsvorschläge zu machen. Als weitere Experten wurden drei wissenschaftlich arbeitende Psychologen mit Arbeitsschwerpunkten in den Bereichen Schriftspracherwerb und Lesekompetenz um eine kritische Betrachtung der Tests gebeten.

Itemrevision. Im Anschluss an die Itemselektion fand eine Überarbeitung der verbleibenden Items auf Basis der Analyseergebnisse sowie der Kritik durch die Schüler, Lehrkräfte und Experten statt. Zudem wurden Distraktoren modifiziert, um zu erreichen, dass erstens die falschen Antwortalternativen etwa gleichwahrscheinlich gewählt werden, zweitens die richtigen Antworten etwa gleichmäßig auf die fünf Antwortpositionen verteilt sind und drittens die Antworten zudem keine Hinweise auf die richtige Lösung enthalten (vgl. Empfehlungen von Bühner, 2011, S. 119). Darüber hinaus wurden die verbleibenden Items durch neue Items ergänzt, wobei darauf geachtet wurde, dass in fehlenden Schwierigkeitsbereichen Items hinzugefügt wurden und am Ende alle Verständnisebenen wiederum gleichmäßig vertreten waren.

13.3.2.2 Ergebnisse

Im folgenden Abschnitt wird zunächst auf deskriptive Statistiken und die Rohwertverteilungen eingegangen. Es folgt die Einführung eines neuen Textes, durch den einer der vorhandenen Texte ersetzt wurde, der sich als zu einfach erwiesen hatte. Weiter

werden die Itemkennwerte und die Ergebnisse der Itemselektion sowie Revisionen der Testentwürfe aufgrund der Zielgruppen- und Expertenbefragung dargestellt.

Deskriptive Statistik und Verteilungen. Die deskriptive Statistik und die Rohwertverteilungen für die einzelnen Texte vor der Itemselektion können Tabelle 3 bzw. Abbildung 8 entnommen werden. Alle Texte bzw. die dazugehörigen Verständnisfragen erwiesen sich als relativ leicht. Nahezu alle Rohwertverteilungen fallen zumindest leicht linksschief aus. Sowohl in der neunten Gymnasialklasse als auch in den neunten Realschulklassen wurde stets von einem oder mehreren Schülern die volle Punktzahl oder nahezu die volle Punktzahl erreicht. Selbst in den sechsten Klassen – auch in der sechsten Hauptschulklasse – erreichte mindestens ein Schüler mehr als drei Viertel der möglichen Punkte. Die minimal erreichte Punktzahl lag in der sechsten Hauptschulklasse bei 15 von 40 maximal möglichen Punkten und bei allen anderen Klassen deutlich darüber. Im Mittel wurden in der sechsten Hauptschulklasse knapp 25 der 40 Punkte erreicht, in der neunten Gymnasialklasse lag der Mittelwert sogar bei 37 von 40 maximal möglichen Punkten.

Tabelle 3. Deskriptive Statistik der Rohwerte für die einzelnen Texte bei der ersten Voruntersuchung zum Subtest TV vor der Itemselektion (jeweils max. 40 Punkte erreichbar).

Text	Textgenre	Schulart	Klasse	*N*	*M*	*SD*	Min-Max
Koboldmakis	expositorisch	RS	6	31	27.81	3.94	20-34
Koboldmakis	expositorisch	RS	9	41	33.05	4.25	19-40
Tiefsee	expositorisch	RS	6	30	29.77	3.47	22-34
Tiefsee	expositorisch	RS	9	30	34.40	3.66	26-39
Tiefsee	expositorisch	GYM	9	14	37.07	2.30	32-40
Gedächtniskunst	narrativ	RS	6	19	31.84	3.70	26-37
Gedächtniskunst	narrativ	RS	9	25	35.68	2.98	29-40
Der geheilte Patient	narrativ	HS	6	18	24.56	4.63	15-33
Der geheilte Patient	narrativ	RS	6	30	28.77	4.59	20-37
Der geheilte Patient	narrativ	RS	9	33	33.76	3.72	21-39

Neuer Text. Der narrative Text „Gedächtniskunst“ und die Fragen dazu waren insgesamt sehr leicht und es erschien bei diesem Text schwierig, eine ausreichende Anzahl an angemessen schweren Items zu generieren. Daher wurde der Text durch den schwereren narrativen Text „Der Gescheitere“ von Johann Peter Habel ersetzt. Für den „neuen“ Text wurden 48 Items erstellt, um für die Itemselektion eine größere Anzahl an Items zur Verfügung zu haben als zuvor bei den anderen Texten. Der Text und die dazugehörigen Fragen wurden in einer sechsten Realschulklasse und zwei neunten Realschulklassen pilotiert. Tabelle 4 zeigt die Stichprobengrößen sowie die deskriptive Statistik, Abbildung 9 zeigt die Rohwertverteilungen. Obwohl die Mittelwerte recht hoch liegen (bei 59 % der max. Punktzahl in der sechsten Klassenstufe und bei 77 %

der max. Punktzahl in der neunten Klassenstufe) erreichte kein Schüler die maximale Punktzahl. Insgesamt wurden zwischen 33 % und 92 % der Items korrekt gelöst.

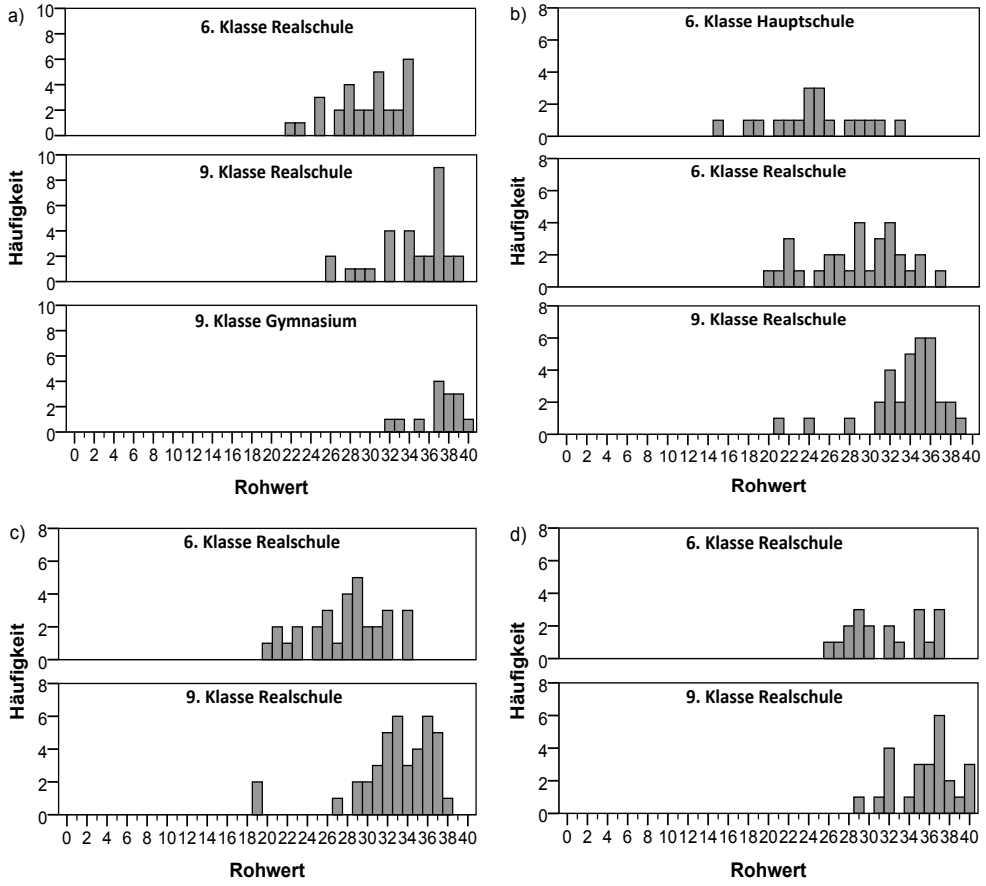


Abbildung 8. Rohwertverteilungen für *Tiefsee* (a), *Der geheilte Patient* (b), *Koboldmakis* (c) und *Gedächtniskunst* (d) bei der ersten Voruntersuchung zum Subtest TV (jeweils max. 40 Punkte erreichbar).

Tabelle 4. Deskriptive Statistik für den narrativen Text *Der Gescheitere* in der ersten Voruntersuchung für den Subtest TV (max. erreichbare Punktzahl: 48).

Text	Textgenre	Schulart	Klasse	*N*	*M*	*SD*	Min	Max
Der Gescheitere	narrativ	RS	6	28	28.14	5.58	16	37
Der Gescheitere	narrativ	RS	9	44	37.16	3.98	26	44

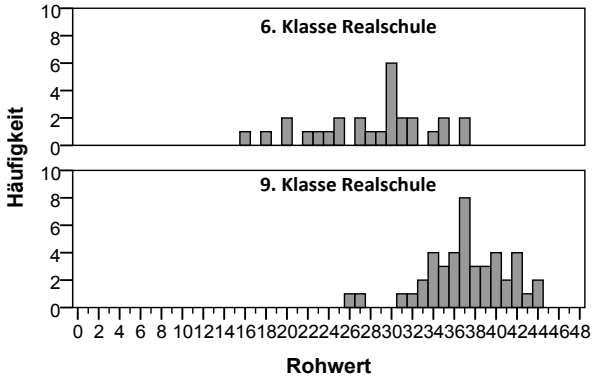


Abbildung 9. Rohwertverteilung für den Text *Der Gescheitere* im Rahmen der ersten Voruntersuchung zum Subtest TV (max. 48 Punkte erreichbar).

Itemkennwerte und Itemselektion. In Anhang A befindet sich für jeden Text eine Tabelle (Tab. 30 bis 33), die für jedes Item die Trennschärfe, die Standardabweichung, den SK sowie den Schwierigkeitsindex angibt. Zusätzlich sind die Verständnisebenen (Kohärenzebene und Repräsentationsform) aufgeführt, da diese zur Beurteilung der inhaltlichen Bedeutung der Items für die vollständige Abbildung des Konstrukts bei der Itemselektion mitberücksichtigt wurden. Da kein Item extrem schwer ($p < .10$) war, wurden lediglich Items mit einem Schwierigkeitsindex von $p > .90$ eliminiert. Diese waren extrem leicht und wurden daher von nahezu allen Schülern korrekt gelöst. Die beibehaltenen Items sind in den Tabellen im Anhang grau markiert. Für die beiden narrativen Texte „Der geheilte Patient“ und „Der Gescheitere“ verblieben je 18 Items im Test, für die expositorischen Texte „Tiefsee“ und „Koboldmakis“ 15 bzw. 17 Items.

Tabelle 5 zeigt die minimale und maximale Trennschärfe sowie die interne Konsistenz für jeden Text nach der erwähnten Itemselektion, allerdings ausschließlich bezüglich der Realschuldaten. Die Trennschärfewerte sind bei den narrativen Texten höher und größtenteils zufriedenstellend, während sie bei den expositorischen Texten zum Teil sehr niedrig ausfallen. Die interne Konsistenz erwies sich ebenfalls bei den narrativen Texten als zufriedenstellend, bei den expositorischen Texten dagegen fielen die Werte relativ niedrig und daher weniger zufriedenstellend aus.

Tabelle 5. Trennschärfe und interne Konsistenz (KR-20) nach der ersten Itemselektion (basierend auf den Realschuldaten).

Text	Itemanzahl	N	$r_{it_{min}}$	$r_{it_{max}}$	KR-20
Der geheilte Patient	18	63	.29	.59	.80
Der Gescheitere	18	72	.31	.56	.84
Tiefsee	15	60	.19	.48	.68
Koboldmakis	17	72	.20	.47	.72

Tabelle 6 zeigt den Anteil korrekt gelöster Items für die einzelnen Texte nach Klassenstufe und Schulart getrennt, und zwar sowohl vor als auch nach der Itemselektion. Es zeigt sich, dass die Items zu den einzelnen Texten nach der Itemselektion insgesamt leichter ausfielen als vorher. Um die Testschwierigkeiten zu erhöhen, standen verschiedene Möglichkeiten zur Verfügung: Die Zahl der Distraktoren hätte z. B. erhöht werden können, was die Ratewahrscheinlichkeit reduziert und die Itemschwierigkeit erhöht hätte. Da jedoch mit fünf Antwortalternativen bereits eine relativ große Auswahl (und somit eine eher geringe Ratewahrscheinlichkeit) gegeben war, schied diese Option aus. Stattdessen wurde durch Umformulierungen und Änderung der Distraktoren versucht, diese gleich „schwierig“ und qualitativ gleich „gut“ zu gestalten. In der ersten Voruntersuchung zeigte sich, dass mehrere Distraktoren von den Probanden überhaupt nicht angekreuzt wurden. An dieser Stelle steckte also am meisten Verbesserungspotenzial.

Tabelle 6. Anteil korrekt gelöster Items für die einzelnen Texte nach Klassenstufe und Schulart getrennt, vor und nach der Itemselektion.

				Itemanzahl		korrekte Lösungen (in %)	
Schulart	Klassenstufe	*N*	Text	vor	nach	vor	nach
HS	6	18	Der geheilte Patient	40	18	56	61
GYM	9	14	Tiefsee	40	15	92	93
RS	6	30	Der geheilte Patient	40	18	68	72
RS	6	28	Der Gescheitere	48	18	55	59
RS	6	30	Tiefsee	40	15	67	74
RS	6	31	Koboldmakis	40	17	63	70
RS	9	33	Der geheilte Patient	40	18	86	84
RS	9	44	Der Gescheitere	48	18	84	77
RS	9	30	Tiefsee	40	15	81	86
RS	9	41	Koboldmakis	40	17	82	83

Zielgruppen- und Expertenbefragung. Für die vorgenannten Modifikationen erwiesen sich die Ergebnisse der Zielgruppen- und Expertenbefragung als sehr nützlich. Durch sie konnten Hinweise z. B. zu missverständlichen Items und Distraktoren gesammelt werden. Zudem flossen Ideen für zusätzliche Verständnisfragen und Distraktoren in die Itemrevisionen mit ein. Darüber hinaus äußerten Schüler der Haupt- und Realschulklassen, dass sie die Aufgaben zum Teil recht schwer, aber noch zumutbar fanden. In den neunten Realschulklassen und in der neunten Gymnasialklasse wurden die Aufgaben als sehr leicht eingeschätzt.

Alle Ergebnisse dieser ersten Voruntersuchung waren jedoch mit Vorsicht zu interpretieren, da die Stichproben sehr klein waren und die Leistungsheterogenität im Vergleich zur eigentlichen Zielpopulation eingeschränkt gewesen sein dürfte, da nur einzelne Klassenstufen und Schularten herausgegriffen wurden.

13.3.3 Zweite Voruntersuchung zum Subtest TV

Ziel der zweiten Voruntersuchung zum Subtest TV war es, die nach der ersten Itemselektion verbliebenen und revidierten Items zu erproben und zu prüfen, ob die Bearbeitung von zwei Texten und jeweils 20 dazugehörigen Verständnisfragen eine angemessene Anforderung an die Schüler der Zielgruppen darstellt. Zudem sollten die Items in Hauptschulen und Gymnasien erprobt werden, da diese Subgruppen der Zielgruppen in der ersten Voruntersuchung nur sehr eingeschränkt untersucht worden waren.

13.3.3.1 Methode

Die Beschreibung der Methode erfolgt analog zur ersten Voruntersuchung. Im Folgenden werden wieder zunächst die Datenerhebung und die Stichprobe beschrieben und anschließend wird auf die Datenanalyse sowie die durchgeführte Schülerbefragung eingegangen.

Datenerhebung. Für die Datenerhebung der zweiten Voruntersuchung wurden entsprechend der Zielsetzung Schulklassen aus Hauptschulen und Gymnasien der Klassenstufen sechs bis neun rekrutiert und jeweils zwei Texte – ein expositorischer Text und ein narrativer Text – kombiniert dargeboten. „Version A“ enthielt die Texte „Koboldmakis“ und „Der geheilte Patient“ und „Version B“ enthielt die Texte „Tiefsee“ und „Der Gescheitere“.

Stichprobe. Die Stichprobe umfasste Hauptschüler der Klassenstufen sechs bis acht und Gymnasiasten der Klassenstufen sechs, sieben und neun. Es war leider nicht gelungen, für beide Schulformen alle Klassenstufen abzudecken, da die neunten Hauptschulklassen kurz vor ihrer Abschlussprüfung standen und die achten Klassen der Gymnasien zum entsprechenden Zeitpunkt an einer bayernweiten Vergleichsstudie teilnahmen. Insgesamt nahmen $N = 201$ Schüler an der zweiten Voruntersuchung zum Subtest TV teil. Aus Tabelle 7 im Ergebnisabschnitt geht zudem die Verteilung der Schüler über die Schularten und Klassenstufen hervor.

Datenanalyse und Schülerbefragung. Die Daten wurden rein deskriptiv analysiert. Auf weitere Auswertungen wurde im Rahmen der zweiten Voruntersuchung zum Subtest TV aufgrund der kleinen Stichproben verzichtet. Auch hier wurden jedoch die Schüler zur Zumutbarkeit der Aufgabenstellung befragt.

13.3.3.2 Ergebnisse

Im folgenden Abschnitt wird zunächst die deskriptive Datenanalyse dargestellt und anschließend werden die Ergebnisse der Schülerbefragung berichtet.

Deskriptive Statistik. Tabelle 7 zeigt Ergebnisse einer deskriptiven Statistik. In der Hauptschule erreichten die Schüler von der sechsten bis zur achten Klasse im Mittel etwa die Hälfte oder etwas weniger als die Hälfte der pro Text maximal erreichbaren Punkte. In den sechsten und siebten Gymnasialklassen erreichten die Schüler etwas mehr Punkte, nämlich die Hälfte bis drei Viertel der pro Text maximal erreichbaren Punkte. Bei Version A wurde bereits ab der siebten Klasse bei beiden Texten von mindestens einem Gymnasiasten die volle Punktzahl erreicht, bei Version B war dies nur in den neunten Gymnasialklassen der Fall. In den neunten Gymnasialklassen lag der Mittelwert bei allen Texten etwa bei drei Vierteln der maximal erreichbaren Punkte. Die niedrigste Punktzahl, die ein Schüler erreichte, war sowohl in den Hauptschulen als auch in den Gymnasien bis zur siebten Klasse sehr niedrig (z. T. deutlich weniger als ein Viertel der möglichen Punkte pro Text). In der Hauptschule lagen die niedrigsten Werte auch in der achten Klasse noch auf ähnlich niedrigem Niveau. Lediglich in den neunten Gymnasialklassen erreichten alle Schüler mindestens die Hälfte der möglichen Punkte pro Text.

Tabelle 7. Stichprobengrößen und deskriptive Statistik für die einzelnen Texte in der zweiten Voruntersuchung zum Subtest TV (maximal 20 Punkte pro Text).

Version A										
			Koboldmakis				*Der geheilte Patient*			
Klasse	Schulart	*N*	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max
6	HS	11	9.18	3.60	5	16	10.72	3.61	5	15
	GYM	13	9.77	2.95	5	14	11.62	4.54	1	17
7	HS	7	9.71	2.21	7	14	9.57	3.69	3	14
	GYM	62	12.50	2.97	4	20	14.65	2.86	5	20
8	HS	10	9.90	3.00	3	13	10.70	2.75	7	16
	GYM	–	–	–	–	–	–	–	–	–
9	HS	–	–	–	–	–	–	–	–	–
	GYM	14	15.21	2.19	11	19	16.86	2.35	13	20
Version B										
			Tiefsee				*Der Gescheitere*			
Klasse	Schulart	*N*	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max
6	HS	12	7.33	3.85	2	14	8.08	2.87	4	13
	GYM	13	11.62	3.23	6	15	10.62	3.28	6	16
7	HS	6	8.67	3.78	4	14	8.67	3.50	4	13
	GYM	30	12.07	3.42	3	17	11.73	3.24	3	17
8	HS	8	8.13	3.44	3	13	8.38	2.56	3	10
	GYM	–	–	–	–	–	–	–	–	–
9	HS	–	–	–	–	–	–	–	–	–
	GYM	14	15.36	2.34	10	20	16.64	1.95	13	20

Schülerbefragung. In der Schülerbefragung beurteilten nur die sechsten und siebten Hauptschulklassen das Schwierigkeitsniveau des Tests als grenzwertig bis zu hoch. Einige dieser Schüler schafften es auch nicht, innerhalb einer Schulstunde alle Fragen zu bearbeiten. Daher wurde entschieden, die Texte für die Klassenstufen sechs und sieben zu kürzen, um beide Texte und den Subtest BLK innerhalb einer Schulstunde vorgeben zu können.

13.3.4 Dritte Voruntersuchung zum Subtest TV

Ziel der dritten Voruntersuchung zum Subtest TV war es, die Texte und Fragen in ihrer endgültigen Kombination an einer größeren Stichprobe von Hauptschülern und Gymnasiasten zu prüfen. Darüber hinaus sollte die strukturelle Schwierigkeit der Texte in ihrer endgültigen Form anhand von Lesbarkeitsindizes bestimmt werden und schließlich sollten über die bisherigen Analysemethoden hinaus auch IRT-Analysen durchgeführt werden.

13.3.4.1 Methode

Im Folgenden wird zunächst auf die Datenerhebung und die Stichprobe eingegangen und anschließend die Bestimmung der strukturellen Textschwierigkeit anhand von Lesbarkeitsindizes beschrieben. Das Vorgehen bei den deskriptiven Analysen, den Verteilungsanalysen, den KTT-Itemanalysen und der Itemselektion sowie bei der Zielgruppen- und Expertenbefragung war analog zur ersten Voruntersuchung und wird daher hier nicht noch einmal dargestellt. Auf die an dieser Stelle erstmals durchgeführten IRT-Analysen wird jedoch eingegangen. Die nach der Itemselektion erneut durchgeführten Itemanalysen und die Reliabilitätsanalysen werden wiederum nicht erläutert, da auch diese bereits bei der ersten Voruntersuchung beschrieben wurden. Die Erstellung der Endversionen wird dagegen dargestellt.

Lesbarkeitsindizes. Die strukturelle Schwierigkeit der einzelnen Texte wurde anhand der Lesbarkeitsindizes „Flesch“ (vgl. Bachmann, 2009) und „LIX“ (vgl. A. Lenhard & Lenhard, 2011) geprüft (s. auch Kap. 7.3). In der vorliegenden Arbeit wurde der von Bachmann (2009) online zur Verfügung gestellte Flesch-Rechner verwendet. Beim Flesch-Index resultiert in der Regel ein Wert zwischen 0 und 100. Bachmann stellt zudem für die Interpretation des Indexes eine Tabelle zur Verfügung, die zum einen eine Schwierigkeitsbeurteilung erlaubt und zum anderen angibt, welche Bildung erforderlich ist, um einen Text des entsprechenden Schwierigkeitsgrades gut verstehen zu können (s. Tab. 8).

Für die Berechnung des LIX wurde der von A. Lenhard und Lenhard (2011) online zur Verfügung gestellte LIX-Rechner verwendet. Der LIX erlaubt eine ungefähre Einschätzung der Textschwierigkeit, wobei sich beim Vergleich verschiedener Textgat-

tungen im Schnitt folgende typische Werte ergeben: < 40 für Kinder- und Jugendliteratur, 40 bis 50 für Belletristik, 50 bis 60 für Sachliteratur und > 60 für Fachliteratur.

Tabelle 8. Beurteilung des Flesch-Indexes bezüglich der Lesbarkeit (Schwierigkeitsbeurteilung und erforderliche Bildung für gutes Verständnis des Textes; vgl. Bachmann, 2009).

Flesch-Index	Beurteilung	erforderliche Bildung
81 - 100	extrem leicht	5. Klasse
71 - 80	sehr leicht	6. bis 8. Klasse
61 - 70	leicht	Abschlussklasse
41 - 60	durchschnittlich	Sekundarschule, Fachoberschule, Berufsschule
31 - 40	etwas schwierig	Mittelschule
21 - 30	schwierig	Abitur
≤ 20	sehr schwierig	Hochschulabschluss

Datenerhebung. Die Texte wurden noch einmal neu kombiniert und folgendermaßen den Klassenstufen zugewiesen: Die Texte „Tiefsee“ und „Der geheilte Patient“, die sich in der ersten Voruntersuchung durchweg als der jeweils leichtere expositorische bzw. narrative Text erwiesen hatten, wurden – wie aufgrund der Ergebnisse der zweiten Voruntersuchung beschlossen – gekürzt und den Klassenstufen sechs und sieben zugewiesen. Die Texte „Koboldmakis“ und „Der Gescheitere“ blieben unverändert und wurden den Klassenstufen acht und neun zugeordnet. Die entsprechenden Items wurden von der zweiten zur dritten Voruntersuchung nicht verändert. Die Beispielaufgabe als Teil der Instruktion wurde jedoch um ein Item gekürzt, um die totale Testzeit nicht unnötig in die Länge zu ziehen. Die Schüler schienen mit dem Aufgabenformat stets ausreichend vertraut zu sein, sodass ein Übungsbeispiel zu genügen schien. Die so entstandenen Subtests TV für LESEN 6-7 und LESEN 8-9 wurden in den entsprechenden Klassenstufen wiederum in Hauptschul- und Gymnasialklassen durchgeführt.

Stichprobe. Insgesamt bearbeiteten 113 Schüler LESEN 6-7 und 189 Schüler LESEN 8-9. Informationen zur Verteilung der Schüler über die Schularten und Klassenstufen können Tabelle 10 im Ergebnisteil entnommen werden.

Wahl des Testmodells. Für die IRT-Skalierung musste zunächst eines der IRT-Modelle ausgewählt werden. Die Entscheidung, der Skalierung das dichotome Rasch-Modell (Rasch, 1960, S. 62ff.) zugrunde zu legen, basierte zum einen auf der Überlegung, dass die Skala für den Subtest TV eindimensional sein sollte (vgl. Kap. 4.3). Zum anderen wurde aus Gründen der Bearbeitungs- und Auswertungsökonomie sowie zur Sicherstellung der Objektivität ein geschlossenes, dichotomes Antwortformat gewählt. Zwar wäre bei diesem Antwortformat, das eine gewisse Lösungswahrscheinlichkeit aufgrund von Raten mit sich bringt, auch die Anwendung eines IRT-Modells denkbar gewesen, das die Ratewahrscheinlichkeit berücksichtigt, jedoch sollte ein Testmodell

u. a. das Einfachheitskriterium erfüllen, d. h. es sollten möglichst wenige Parameter geschätzt werden müssen (J. Rost, 2004, S. 330). Es erscheint recht gewagt, aus der geringen Information, die ein dichotomes Antwortformat enthält, mehr als zwei Parameter zu schätzen. Auch DeMars (2010, S. 29) spricht sich dafür aus, ein einfaches Modell, das genau geschätzt werden kann, gegenüber einem komplexeren Modell zu bevorzugen, das die Daten zwar möglicherweise besser repräsentiert, das jedoch nur schlecht geschätzt wird.

Aus diesen Gründen wurde anstatt der Verwendung eines IRT-Modells, das die Ratewahrscheinlichkeit berücksichtigt, versucht diese mithilfe einer relativ großen Anzahl an qualitativ hochwertigen Distraktoren gering zu halten. Hierfür wurden bereits zuvor mehrere Distraktorenanalysen durchgeführt. Bei der dritten Voruntersuchung wurden darüber hinaus zur Einschätzung, ob die Schüler die Items gewissenhaft bearbeiteten oder die Antwort nur rieten, die empirischen Daten mit hypothetischen Punktzahlen verglichen, welche bei rein zufälligem Raten zu erwarten gewesen wären. Hätten Schüler ohne jegliches Vorwissen und ohne den Text gelesen zu haben, bei allen Items die Antwort geraten, würde man für jeden dieser Texte mit 20 Items, fünf Antwortalternativen und jeweils einer richtigen Lösung erwarten, dass sie etwa vier Punkte erreichen – insgesamt also acht Punkte für den Subtest TV.

Auf eine Prüfung der globalen Modellpassung anhand des Vergleichs des gewählten Modells mit anderen Modellen wurde verzichtet. Zwar gibt es inzwischen einige Ansätze, wie dies statistisch geprüft werden könnte, jedoch bedarf es noch weiterer Forschung, um entscheiden zu können, welcher Test für welche Modelle angemessen ist (DeMars, 2010, S. 59). DeMars (2010, S. 59) empfiehlt, für die Entscheidung für ein Modell statt eines globalen Modelltests den Inhalt und die Zielsetzung des zu entwickelnden Tests als Kriterien heranzuziehen. Ein Vergleich mehrerer Modelle sei nur dann notwendig, wenn aufgrund theoretischer Überlegungen mehrere Modelle in Frage kämen. Auch Bühner (2011, S. 546) empfiehlt auf die Berechnung des Likelihood-Quotienten-Tests gegen das saturierte Modell sowie den Pearson-χ^2-Test zu verzichten, da die Voraussetzungen für diese Verfahren praktisch nie erfüllt sind. Theoretisch könnte man das Bootstrap-Verfahren verwenden, um die Verletzung der Voraussetzungen zu umgehen, jedoch wird von (Bühner, 2011, S. 546) empfohlen, auch auf dieses Verfahren zu verzichten, da es für die Identifikation von Modellverletzungen wenig teststark ist. Aus den genannten Gründen werden globale Modelltests bzw. Vergleiche der Eignung mehrerer Modelle bisher auch eher selten berichtet und meist beschränkt man sich auf die Prüfung des Item-Fits (DeMars, 2010, S. 57). Daher wurde auch in der vorliegenden Arbeit zunächst der Item-Fit betrachtet und entsprechende Kennwerte wurden zur Itemselektion hinzugezogen. Im Anschluss an die Itemselektion erfolgte eine Prüfung der Eindimensionalität der Skalen, der lokalen Unabhängigkeit der Items und der Personenhomogenität, da diese, wie in Kapitel 5.3.1.2 erläutert, Voraussetzungen für die Gültigkeit des Rasch-Modells sind.

Item-Fit. Zur Prüfung des Item-Fits wurden die gewichteten Abweichungsquadrate (weighted Mean Squares, wMNSQ) zwischen der empirischen und der theoretisch erwarteten Verteilung für jedes Item betrachtet. Jedes quadrierte Residuum wird dabei mit seiner Varianz gewichtet, um zu verhindern, dass sehr vereinzelte „Ausreißer“ in den Daten (extrem unerwartete Reaktionen auf ein Item) schon dazu führen, dass ein Misfit des Items zum Modell angezeigt wird. Die Varianz ist umso größer, je besser Personen- und Itemschätzer übereinstimmen und kleiner, wenn der Schätzer der Itemschwierigkeit ober- bzw. unterhalb der Personenfähigkeit liegt. Durch die Gewichtung verlieren einzelne extreme Abweichungen daher an Bedeutung. Ein wMNSQ-Wert von 1 bedeutet eine optimale Passung des Items. Ein wMNSQ-Wert < 1 bedeutet, dass die empirische Verteilung der Lösungshäufigkeiten weniger streut als erwartet, d. h. die beobachtete Item Characteristic Curve (ICC) ist für dieses Item steiler als erwartet. Das Item mit wMNSQ-Wert < 1 weist also – verglichen mit allen anderen Items – eine überdurchschnittlich hohe Trennschärfe auf. Werte von wMNSQ > 1 deuten auf einen flacheren Verlauf der beobachteten ICC im Vergleich zur theoretisch erwarteten ICC hin. Für die Gültigkeit des Rasch-Modells ist es notwendig, dass sich die Trennschärfen nicht signifikant voneinander unterscheiden, d. h. dass es keine signifikanten Abweichungen der wMNSQ-Werte von 1 gibt.

Die wMNSQ-Werte wurden in der vorliegenden Arbeit als zusätzliches Itemselektionskriterium herangezogen. Häufig werden wMNSQ-Werte zwischen 0.75 und 1.33 als akzeptabel bezeichnet (M. Adams & Guillemin, 1996; Bond & Fox, 2001). Daher sollten auch in der vorliegenden Arbeit, wie z. B. von Wilson (2005, S. 129) empfohlen, Items mit wMNSQ < 0.75 bzw. wMNSQ > 1.33 eliminiert werden. Darüber hinaus wurde für jedes Item ein *T*-Wert berechnet, der angibt, ob die Abweichung zwischen der empirischen und der theoretisch erwarteten Verteilung signifikant ist. Werte mit einem Betrag von $T > |2|$ gelten als problematisch und sollten daher ebenfalls eliminiert werden.

Eindimensionalität. Zur Prüfung der Eindimensionalität der Skalen wurde nach der Itemselektion wie von DeMars (2010, S. 40f.) beschrieben ein *Screetest* nach Cattell durchgeführt. Dabei werden die anhand einer exploratorischen Faktorenanalyse (EFA) ermittelten Eigenwerte in einem Screeplot dargestellt. Die Anzahl der der Skala zugrunde liegenden Dimensionen bzw. die Anzahl der zu extrahierenden Faktoren entspricht der Anzahl der Eigenwerte, die im Screeplot von links nach rechts betrachtet vor einem markanten „Knick“ liegen.

Die Ermittlung der Eigenwerte gemäß EFA wurde mit der Software Mplus durchgeführt, da diese die Möglichkeit bietet, tetrachorische Korrelationen zu verwenden. Tetrachorische Korrelationen sind dann angemessen, wenn die zugrunde liegende Variable kontinuierlich und normalverteilt ist, während die beobachtete Variable dichotom ist (DeMars, 2010, S. 40). Bei dichotomen Daten besteht ansonsten die Gefahr, dass die Items aufgrund ihrer Schwierigkeitsniveaus anstatt aufgrund inhaltlicher Ge-

meinsamkeiten zusammengefasst werden. Die Parameterschätzung erfolgte über die WLSMV-Methode, eine gewichtete Kleinste-Quadrate- (Weighted Least Square-) Methode, bei der durch eine entsprechende Gewichtung die Mittelwerte (M) und Varianzen (V) berücksichtigt werden und die von L. K. Muthén für die Durchführung einer EFA mit dichotomen Daten empfohlen wird (http://www.statmodel.com/discussion/messages/23/121.html?1321038861, Zugriff am 19.10.2013).

Da der Screetest jedoch häufig als subjektiv interpretierbar kritisiert wird und dessen Anwendung auch unter Experten zu sehr unterschiedlichen Interpretationen führen kann, sollten weitere Analysen für die Entscheidung über die Anzahl der zugrunde liegenden Dimensionen herangezogen werden (vgl. Crawford & Koopman, 1979; Streiner, 1998). Die u. a. von Bühner (2011, S. 328) und O'Connor (2000) empfohlenen Methoden der Wahl zur Bestimmung der Anzahl zu extrahierender Faktoren – die Parallelanalyse nach Horn und Velicer's Minimum Average Partial- (MAP-) Test (Velicer, 1976; Velicer, Eaton & Fava, 2000) – stehen jedoch bei Mplus nicht zur Verfügung und wurden daher mit SPSS und somit nicht mit tetrachorischen Korrelationen durchgeführt. Weng und Cheng (2012) fanden jedoch, dass zumindest für die Parallelanalyse phi-Korrelationen zur Bestimmung der Anzahl zu extrahierenden Faktoren geeigneter sind als tetrachorische Korrelationen. Der Wert von phi-Korrelationen entspricht numerisch dem der Pearson-Korrelation r, wenn diese bei dichotomen Daten angewendet wird (Bortz & Schuster, 2010, S. 174) und somit jenen Korrelationen, die SPSS bei der Parallelanalyse verwendet.

Bei der *Parallelanalyse* wird der empirische Eigenwertverlauf einem Eigenwertverlauf gegenübergestellt, der aus mehreren Hauptkomponentenanalysen mit Zufallswerten resultiert. Ein Faktor ist dann bedeutsam, wenn der empirische Wert über dem 95 %-Perzentil der Eigenwerte für die Komponente liegt, die aus Zufallszahlen generiert wurde. Hierfür wurde auf die von O'Connor (2000) angegebene SPSS-Syntax zurückgegriffen. Es wurden Parallelanalysen für Hauptachsenanalysen durchgeführt und 1000 Zufallsdatendateien mit Item- und Personenzahlen erzeugt, welche den empirischen Daten des jeweiligen Tests entsprechen (vgl. Bühner, 2011, S. 323ff.).

Beim *MAP-Test* wird anhand einer Hauptkomponentenanalyse die erste Hauptkomponente extrahiert und diese anschließend aus der empirischen Korrelationsmatrix auspartialisiert, wodurch man eine Residualmatrix erhält. Im nächsten Schritt wird die mittlere quadrierte Partialkorrelation der Residualmatrix gebildet. Diese drei Berechnungsschritte werden so oft wiederholt, bis sich die mittlere quadrierte Partialkorrelation nicht mehr reduzieren lässt. Es werden also so lange Komponenten aus der Korrelationsmatrix herauspartialisiert bis keine systematischen Zusammenhänge der Items in der verbleibenden Korrelationsmatrix mehr bestehen. Die Anzahl der zu diesem Zeitpunkt auspartialisierten Hauptkomponenten entspricht der Anzahl zu extrahierender Faktoren. Auch für dieses Vorgehen steht eine SPSS-Syntax von O'Connor (2000) zur Verfügung. Als Kriterium wird das Ergebnis der revidierten Version des MAP-Tests aus dem Jahr 2000 (Velicer et al., 2000) herangezogen. Diese quadriert

nicht die partiellen Korrelationen, sondern bildet deren vierte Potenz. Beim MAP-Test wird zur Faktorenextraktion die Hauptkomponentenanalyse verwendet, die auf diese Weise ermittelte Komponentenanzahl wird jedoch auch im Rahmen der Hauptachsenfaktorenanalyse verwendet (Bühner, 2011, S. 325).

Darüber hinaus existieren weitere Methoden der Faktorenextraktion. Das *Eigenwertkriterium größer eins-* und das *Kaiser-Guttmann-Kriterium* empfehlen z. B. alle Faktoren mit einem Eigenwert > 1 zu extrahieren. Sie führen jedoch häufig zur Über- und in seltenen Fällen zur Unterschätzung der tatsächlichen Faktorenanzahl (W. R. Zwick & Velicer, 1986) und werden daher nicht mehr empfohlen (z. B. Bühner, 2011; O'Connor, 2000). Diese Kriterien werden daher in der vorliegenden Arbeit nicht berücksichtigt.

Vor der Durchführung der EFA war zu prüfen, ob die *Voraussetzungen* dafür erfüllt sind (vgl. Bühner, 2011, S. 342ff.). So ist eine Stichprobengröße von $N \geq 60$ erforderlich. Weiter sollten die Itemreliabilitäten mindestens $r = .60$ betragen, um die Ladungen ausreichend genau schätzen zu können. Bei größeren Stichproben ($N > 100$) sind auch geringfügig niedrigere Itemreliabilitäten akzeptabel (vgl. MacCallum, Widaman, Zhang & Hong, 1999). Als Mindestschätzung der Reliabilität eines Items kann die höchste Korrelation des Items mit einem anderen Item dienen (Bühner, 2011, S. 344). Die weiteren Voraussetzungsprüfungen wurden mit SPSS durchgeführt, da diese in Mplus nicht verfügbar sind. Mithilfe des Kaiser-Meyer-Olkin- (KMO-) Koeffizienten, wurde geprüft, ob substanzielle Korrelationen in der Korrelationsmatrix vorliegen. Bühner (2011, S. 348) nennt als Mindestwert für die Durchführung einer EFA einen Wert von KMO $= .50$. Weiter wurde der Measure of Sample Adequacy- (MSA-) Koeffizient bestimmt. Dieser prüft jedes einzelne Item auf die Eignung für eine EFA. Es wird also die Korrelation eines einzelnen Items mit den restlichen Items betrachtet. Auch dieser Koeffizient sollte mindestens bei MSA $= .50$ liegen. Schließlich wurde noch der Bartlett-Test durchgeführt, der die globale Nullhypothese prüft, die besagt, dass alle Korrelationen der Matrix gleich Null sind. Kann die Nullhypothese abgelehnt werden, ist die Matrix faktorisierbar. Hier ist jedoch zu beachten, dass der Test nur eine Minimalbedingung darstellt, da er bei einer ausreichend großen Stichprobe sehr schnell signifikant wird (Bühner, 2011, S. 348).

Für die Bewertung der *Faktorladungen* gibt es verschiedene Kriterien, wobei zu beachten ist, dass es sich bei der EFA um ein exploratives Verfahren handelt, weshalb die Kriterien zur Bedeutsamkeit von Faktorladungen nicht zu streng ausgelegt werden sollten (Bortz & Schuster, 2010, S. 422). Eine Faustregel von Gorsuch (1983, S. 2010) besagt, dass Faktorladungen mindestens $\lambda = .30$ betragen sollten. Kline (2002, S. 52f.) bestätigt dies und fügt hinzu, dass nicht nur die statistische, sondern auch die praktische Bedeutsamkeit von Faktorladungen zu berücksichtigen ist.

Lokale Unabhängigkeit. Die Prüfung der lokalen Unabhängigkeit erfolgte wie von DeMars (2010, S. 48ff) beschrieben anhand des Q_3-Tests nach Yen (1984). Dieser Test basiert auf der Idee, dass Items eines Tests bei lokaler Unabhängigkeit nach Kontrol-

le der Personenfähigkeit nicht mehr korrelieren sollten. Hierfür werden zunächst die Item- und Personenparameter geschätzt. Basierend auf dieser Schätzung der Itemparameter werden Residuen für die Reaktionen aller Personen berechnet. Diese Residuen ergeben sich aus der Differenz zwischen der vorhergesagten Itemantwort und der tatsächlich beobachteten Itemantwort jeder Person für jedes Item. Bei dichotomen Items ergibt sich das Residuum e aus $e = P(\Theta) - u$. Dabei stellt $P(\Theta)$ die Wahrscheinlichkeit einer richtigen Lösung (gegeben die geschätzten Item- und Personenparameter) dar und u die tatsächlich beobachtete Itemantwort einer Person auf das dichotome Item (0 = falsche Lösung oder 1 = korrekte Lösung). Anschließend wird die Korrelationsmatrix der Residuen betrachtet, um herauszufinden, ob es Itempaare mit hohen Residual-Korrelationen gibt. Da bei großen Stichproben die Signifikanz von Korrelationen wenig aussagekräftig ist, wird von Yen ein Grenzwert von $Q_3 = .20$ vorgeschlagen. Diesen Grenzwert von Q_3 sollten die Residual-Korrelationen zweier Items nicht überschreiten, damit lokale Unabhängigkeit der Items angenommen werden kann.

Personenhomogenität. Eine Voraussetzung für spezifisch objektive Vergleiche besteht darin, dass ein Test bei jeder beliebigen Substichprobe dasselbe Merkmal misst, d. h. dass die bei Geltung des Rasch-Modells unabhängig von den Personenparametern schätzbaren Itemparameter in jeder Substichprobe gleich sind (Personenhomogenität). Um dies zu prüfen, wurden grafische Modelltests sowie Andersen-Tests durchgeführt.

Die *Grafischen Modelltests* sind keine Modelltests im eigentlichen Sinne, sondern rein deskriptive Analysen. Dabei werden die Gesamtstichproben jeweils in zwei Substichproben aufgeteilt und für beide Substichproben die Itemparameter separat geschätzt. Anschließend werden die Itemparameter in Streudiagramme eingetragen, wobei die Daten der einen Substichprobe an der x-Achse, die der anderen Substichprobe an der y-Achse aufgetragen werden. Liegen die Itemparameter nun auf einer Winkelhalbierenden, bedeutet dies, dass sie für beide Substichproben übereinstimmen, was für Personenhomogenität spricht. Wie in der Literatur üblich (vgl. z. B. Bühner, 2011; J. Rost, 2004) wurde in der vorliegenden Arbeit als Teilungskriterium der Median des Rohwerts für den Subtest TV herangezogen. Es gäbe aber zahlreiche weitere Möglichkeiten, die Stichprobe zu teilen (z. B. Geschlecht, Alter). Aus dem Funktionieren des Tests für zwei Substichproben kann dabei nicht auf das Funktionieren für alle möglichen anderen Stichprobenaufteilungen geschlossen werden. Theoretisch müssten somit alle möglichen Teilungskriterien geprüft werden, dies ist jedoch praktisch nicht möglich. Aus diesem Grund wurde lediglich die Teilung am Median des Rohwerts getestet, was einem verbreiteten und akzeptierten Vorgehen entspricht.

Um zu prüfen, ob die Abweichungen der Items vom Modell signifikant ausfallen, kann der bedingte Likelihood-Quotienten-Test – auch *Andersen-Test* – durchgeführt werden (Andersen, 1972). Dieser ist generell sehr sensitiv gegenüber Modellverlet-

zungen (vgl. Bühner, 2011, S. 548). Auch hier wird die Stichprobe in mehrere Teile aufgeteilt und die gemeinsame Likelihood der gesamten Stichprobe (Zähler) mit den Likelihoods der Substichproben (Nenner) verglichen (vgl. Strobl, 2010, S. 43). Ergibt sich ein LQ-Wert von 1 oder nahe 1, stimmen die gemeinsame Likelihood und die Likelihoods der Substichproben überein, was für Personenhomogenität spricht. Ob die Abweichung von 1 signifikant ist, kann statistisch geprüft werden, denn die Teststatistik $T = -2ln \cdot LQ$ ist χ^2-verteilt. Somit wäre bei gegebener Personenhomogenität $T = 0$, und bei einer Verletzung der Personenhomogenität $T > 0$ (da LQ < 1). Große *T*-Werte sprechen somit für eine signifikante Modellverletzung. Das Signifikanzniveau wurde für den Test auf $\alpha = .20$ festgelegt, da die Nullhypothese bestätigt und somit der β-Fehler klein gehalten werden sollte. Der auf diese Weise empirisch ermittelte χ^2-Wert wurde mit einem kritischen χ^2-Wert mit der entsprechenden Anzahl an Freiheitsgraden verglichen. Eigentlich müsste dieser Test – wie der grafische Modelltest – für alle möglichen Teilungskriterien durchgeführt werden, was auch hier praktisch nicht möglich ist. Es wird daher wieder der Median als übliches Teilungskriterium herangezogen.

Erstellung der Endversionen. Abschließend erfolgte die Erstellung der Endversionen. Basierend auf den Daten der dritten Voruntersuchung wurden schließlich die für die Endversion des Subtests TV ausgewählten Items zu jedem Text aufsteigend nach Schwierigkeit angeordnet. Leicht zu lösende Items wurden an den Anfang gestellt, um den Einstieg in die Bearbeitung zu erleichtern und schwache Leser nicht gleich zu entmutigen. Zudem kann – sollte ein Schüler in der vorgegebenen Zeit nicht mit allen Aufgaben fertig werden – bei Fehlern in den ersten Aufgaben vermutet werden, dass er die schwersten Aufgaben am Ende ohnehin nicht hätte lösen können. Dieser Fall sollte jedoch äußerst selten vorkommen, da die Zeitbegrenzung derart großzügig gewählt wurde, dass es in der dritten Voruntersuchung allen Schülern gelang, alle Aufgaben zu bearbeiten. Aufgrund der Annahme der Unabhängigkeit der Items voneinander sollte es keine Reihenfolgeeffekte oder Effekte der Lösung bzw. Nichtlösung eines Items auf die Beantwortung weiterer Items geben.

13.3.4.2 Ergebnisse

Im Folgenden werden zunächst die Werte der Lesbarkeitsindizes betrachtet sowie deskriptive Werte und die Rohwertverteilungen vor der Itemselektion. Anschließend wird geprüft, ob die Ergebniswerte der Schüler über den hypothetischen Werten liegen, die sie bei reinem Raten erreicht hätten. Es folgt die Darstellung der Ergebnisse der Item- und Distraktorenanalyse sowie der Itemselektion. Zuletzt wird auf die Ergebnisse der Überprüfung der Rasch-Modell-Voraussetzungen Eindimensionalität, lokale Unabhängigkeit und Personenhomogenität eingegangen.

Lesbarkeitsindizes. Die Lesbarkeitsindizes für die Texte von LESEN 6-7 und LESEN 8-9 sind in Tabelle 9 ersichtlich. Sowohl der Flesch-Index als auch der LIX kommen zu dem Ergebnis, dass die expositorischen Texte beider Tests strukturell schwieriger sind als die narrativen Texte. Bei LESEN 6-7 fällt der expositorische Text „Tiefsee" dem Flesch-Index zufolge „etwas schwierig" aus und erfordert für gutes Verständnis etwa einen Mittelschulabschluss. Der LIX stuft den Text als „mittelschwer" ein und schreibt ihm die strukturelle Schwierigkeit eines durchschnittlichen Sachliteraturtextes zu. Der narrative Text von LESEN 6-7 („Der geheilte Patient") wird vom Flesch-Index als „sehr leicht" bewertet und erfordert für ein gutes Verständnis demnach das Verständnisniveau eines Sechst- bis Achtklässlers. Der LIX kommt zu einem sehr ähnlichen Ergebnis. Ihm zufolge ist der Text ebenfalls sehr leicht und entspricht dem strukturellen Schwierigkeitsniveau durchschnittlicher Kinder- und Jugendliteratur.

Bei LESEN 8-9 stuft der Flesch-Index den expositorischen Text „Koboldmakis" als „etwas schwierig" ein. Er erfordert demnach für ein gutes Verständnis einen Mittelschulabschluss. Dem LIX zufolge ist der Text mittelschwer und weist das strukturelle Schwierigkeitsniveau durchschnittlicher Belletristik auf. Der narrative Text „Der Gescheitere" wird vom Flesch-Index als strukturell „sehr leicht" bewertet und erfordert das Verständnisniveau von Sechst- bis Achtklässlern. Der LIX schreibt dem Text ebenfalls ein niedriges strukturelles Schwierigkeitsniveau zu, das etwa dem durchschnittlicher Kinder- und Jugendlichenliteratur entspricht.

Insgesamt können die expositorischen Texte für die Zielgruppen von LESEN 6-7 und LESEN 8-9 als recht anspruchsvoll angesehen werden, während die narrativen Texte als strukturell einfach eingestuft werden und für die entsprechenden Zielgruppen gut verständlich sein sollten.

Tabelle 9. Textmerkmale sowie Flesch- und LIX-Indizes der Texte von LESEN 6-7 und LESEN 8-9.

LESEN 6-7								
	Anzahl Sätze	Satzlänge	Anzahl Wörter ges.	Anzahl Wörter versch.	Anzahl Silben	% lange Wörter	Flesch	LIX
Tiefsee	32	18	555	309	990	32.7	38	50.7
Der geh. Patient	47	18	705	342	949	14.6	78	32.6
LESEN 8-9								
	Anzahl Sätze	Satzlänge	Anzahl Wörter ges.	Anzahl Wörter versch.	Anzahl Silben	% lange Wörter	Flesch	LIX
Koboldmakis	44	16	708	389	1 280	32.9	39	48.9
Der Gescheitere	61	18	897	349	1 273	18.6	72	36.5

Deskriptive Statistik und Verteilungen vor der Itemselektion. Tabelle 10 zeigt deskriptive Statistiken sowohl für die einzelnen Texte als auch für beide Texte zusammen. Während bei LESEN 6-7 kein Schüler die volle Punktzahl für einen der Texte erreichte, kam dies beim expositorischen Text von LESEN 8-9 bereits in der achten Klasse bei einigen Schülern vor. In der neunten Klasse wurde sowohl beim expositorischen Text als auch beim narrativen Text die volle Punktzahl erreicht, wenngleich es keinem Schüler gelang bei beiden Texten die volle Punktzahl zu erreichen.

Insgesamt erreichten bei LESEN 6-7 die Hauptschüler sowohl in der sechsten als auch in der siebten Klasse über beide Texte im Mittel etwas weniger als die Hälfte der möglichen Punkte. Die Gymnasiasten der sechsten Klasse erreichten etwas mehr als 60 % der möglichen Punkte und die Gymnasiasten der siebten Klasse etwa drei Viertel der möglichen Punkte.

Tabelle 10. Deskriptive Statistik für die einzelnen Texte und für beide Texte zusammen in der dritten Voruntersuchung zum Subtest TV vor der Itemselektion (max. 20 Punkte pro Text, 40 Punkte insgesamt).

LESEN 6-7								
	Schul-		expositorischer Text		narrativer Text		Beide Texte	
Kl.	art	*N*	*M (SD)*	*Min-Max*	*M (SD)*	*Min-Max*	*M (SD)*	*Min-Max*
6	HS	33	7.79 (3.29)	1-14	11.00 (3.55)	4-19	18.79 (6.10)	5-33
	GYM	25	11.48 (3.54)	3-17	13.80 (3.08)	8-19	25.28 (5.70)	12-33
7	HS	17	8.35 (3.37)	4-15	10.53 (2.74)	7-15	18.88 (5.09)	11-27
	GYM	39	13.90 (2.21)	7-18	15.74 (2.12)	11-19	29.64 (3.85)	18-36
LESEN 8-9								
	Schul-		expositorischer Text		narrativer Text		Beide Texte	
Kl.	art	*N*	*M (SD)*	*Min-Max*	*M (SD)*	*Min-Max*	*M (SD)*	*Min-Max*
8	HS	21	8.52 (3.34)	3-16	8.57 (3.17)	3-14	17.10 (5.78)	6-26
	GYM	36	15.11 (3.27)	8-20	14.75 (2.95)	8-19	29.86 (5.72)	18-39
9	HS	80	10.26 (3.45)	3-18	11.06 (4.04)	3-19	21.33 (6.83)	6-35
	GYM	52	15.85 (3.01)	8-20	16.06 (3.33)	2-20	31.90 (5.63)	10-39

Bei LESEN 8-9 erreichten die Hauptschüler der achten Klasse im Mittel nur knapp 43 % der möglichen Punkte, die Hauptschüler der neunten Klasse dagegen erreichten im Mittel etwas mehr als die Hälfte der möglichen Punkte. Die Gymnasiasten der achten Klasse erreichten im Mittel etwa drei Viertel der möglichen Punkte und die Gymnasiasten der neunten Klasse im Mittel knapp 80 % der möglichen Punkte. Bei der Interpretation der Ergebnisse ist jedoch zu berücksichtigen, dass die Stichproben jeweils sehr klein waren und es sich zum Teil um einzelne Klassen handelte.

Abbildung 10 zeigt die Rohwertverteilungen für den Subtest TV (beide Texte) sowohl für LESEN 6-7 als auch für LESEN 8-9 nach Klassenstufen und Schularten getrennt. Es zeigt sich, dass es in keiner Klassenstufe und keiner Schulart zu einem Bodeneffekt kommt. In der siebten, achten und neunten Klasse fallen die Verteilungen bei den Gymnasiasten linksschief aus, und insbesondere in der neunten Klasse wur-

den insgesamt recht hohe Werte erreicht. Von einem Deckeneffekt kann jedoch auch in diesem Fall nicht gesprochen werden, da kein Schüler die volle Punktzahl erreichte. Insgesamt scheinen die Items somit auch in den Randbereichen des Leistungsspektrums noch differenzieren zu können.

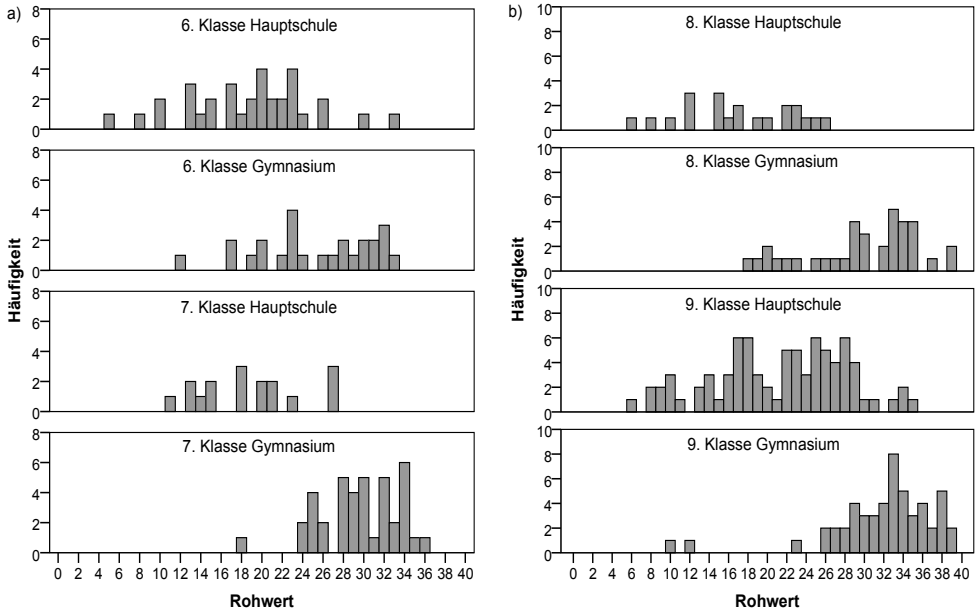


Abbildung 10. Rohwertverteilungen für den Subtest TV für LESEN 6-7 (a) und LESEN 8-9 (b); jeweils max. 40 Punkte erreichbar.

Ratekontrolle. Der Vergleich der Punktzahlen, die bei Raten per Zufall zu erwarten waren, mit den empirischen Daten zeigt, dass bei LESEN 6-7 nur ein Schüler der sechsten Hauptschulklasse weniger als hypothetisch erratene 8 Punkte – nämlich 5 Punkte – erreichte. Bei LESEN 8-9 blieben zwei Schüler unter dem bei zufälligem Ankreuzen erwarteten Wert. Es handelte sich um einen Schüler der achten und einen Schüler der neunten Hauptschulklasse (beide erreichten 6 Punkte). Somit ist anzunehmen, dass die meisten Schüler nicht willkürlich ankreuzten.

Distraktoren- und Itemanalysen sowie Itemselektion. Die Qualität der Distraktoren wurde in der dritten Voruntersuchung bei beiden Tests als zufriedenstellend beurteilt. In den Tabellen 34 und 35 im Anhang finden sich die KTT- und IRT-Kennwerte für alle Items. In den Itemanalysen erfüllten – wie im Folgenden noch ausführlicher beschrieben wird – einzelne Items nicht die Kriterien und wurden daher eliminiert. Sie sind in den Tabellen entsprechend markiert.

Bei LESEN 6-7 erwies sich Item 8 als problematisch. Dieses Item wurde auch von den Deutschlehrkräften als uneindeutig kritisiert. Aus diesem Grund wurde das Item eliminiert. Weiter erfüllten für jeden Text jeweils drei Items nicht die KTT-Kriterien. Für den expositorischen Text waren dies die Items 4, 19 und 20. Die Items 4 und 19 wiesen eine sehr geringe Trennschärfe von $r_{it} < .20$ auf und Item 20 erwies sich als sehr schwer ($p_i < .10$). Die Items 19 und 20 wurden daher eliminiert. Item 4 wurde dagegen weiterhin im Test behalten, da es als inhaltlich wichtig erachtet wurde. Zwar fiel für dieses Item die Trennschärfe gering aus, sie weicht jedoch signifikant von Null ab und zudem liegt der Selektionskennwert über $SK = .20$, sodass ein Beibehalten des Items vertretbar war. Somit wurden insgesamt beim expositorischen Text die Items 8, 19 und 20 eliminiert. Beim narrativen Text wiesen die Items 21, 25 und 32 eine sehr niedrige Trennschärfe von $r_{it} < .20$ auf, die auch für keines der Items das Signifikanzniveau erreicht. Die Items wurden daher alle drei eliminiert. Bei der Rasch-Analyse lagen die wMNSQ-Werte für alle Items zwischen .75 und 1.33, zwei Items des narrativen Textes (Item 25 und Item 32) erfüllten jedoch das Kriterium $T < |2|$ nicht. Es handelt sich dabei um zwei Items, die auch die KTT-Kriterien nicht erfüllten und daher ohnehin eliminiert wurden. Bei LESEN 6-7 wurden also insgesamt drei Items pro Text ausgeschlossen, jeweils 17 wurden beibehalten. In einer erneuten Rasch-Analyse mit den verbleibenden 34 Items wies keines mehr einen *T*-Wert von $T > |2|$ auf. Dieses Ergebnis galt es jedoch anhand einer neuen Stichprobe zu validieren. Dies geschah auf Basis der Normdaten (s. Kap. 15).

Bei LESEN 8-9 wurde ebenfalls ein Item (Item 11) aufgrund einer ungünstigen Formulierung eliminiert, auf die sowohl Schüler als auch Lehrkräfte hingewiesen hatten. Jeweils ein Item pro Text erfüllte nicht die gewünschten KTT-Itemgütekriterien: im expositorischen Text erwies sich die Trennschärfe von Item 8 als nicht signifikant. Da es aber eines der wenigen Items ist, die sich auf Informationen beziehen, die wörtlich im Text zu finden sind, wurde dieses Item aus inhaltlichen Gründen beibehalten. Item 39 des narrativen Textes wies ebenfalls eine nicht signifikante Trennschärfe auf, zusätzlich war hier die richtige Lösung jedoch uneindeutig, sodass dieses Item entfernt wurde. Bei der Rasch-Analyse lagen die wMNSQ-Werte aller Items im gewünschten Bereich von 0.75 bis 1.33. Allerdings wiesen vier Items einen *T*-Wert von $T > |2|$ auf. Dabei handelt es sich wieder um die beiden bereits genannten Items 8 und 39, hinzu kommen noch die Items 13 und 32. Da es sich bei Item 13 um eines der wenigen Items zur Metaebene und das einzige Item zur Autorenintention handelt, wurde es aus inhaltlichen Gründen dennoch beibehalten. Item 32 stellt ebenfalls ein wichtiges Item dar, da es eine emotionale Inferenz erfordert, was für das Verständnis eines narrativen Textes als besonders wichtig erachtet wird. Auch Item 32 wurde also beibehalten. Somit wurde bei LESEN 8-9 ein Item pro Text (Item 11 und Item 39) ausgeschlossen und es verblieben jeweils 19 Items pro Text und 38 Items insgesamt. In einer erneuten Rasch-Analyse mit den verbleibenden Items wiesen drei Items einen *T*-Wert $T > |2|$ auf. Allerdings war bei LESEN 8-9 auch die Stichprobe deutlich größer als bei

bei LESEN 6-7, was die Wahrscheinlichkeit signifikanter Ergebnisse erhöht. Auch bei LESEN 8-9 mussten die Ergebnisse für die finale Version an einer neuen Stichprobe validiert werden, was ebenfalls anhand der Normdaten geschah (s. Kap. 15).

Tabelle 11 zeigt verschiedene statistische Auswertungen der Itemkennwerte für beide Tests vor und nach der Itemselektion. Es ergaben sich zufriedenstellende mittlere Trennschärfen ($M_{r_{it}}$), wobei die kleinsten Trennschärfen trotz Verbesserung im Vergleich zu vor der Revision immer noch für beide Tests sehr niedrig ausfallen. Die mittleren Schwierigkeiten (M_p) sowie die internen Konsistenzen waren schon vor der Revision für beide Tests gut bis sehr gut und sind es auch noch danach. Die wMNSQ-Werte konnten durch die Revision für beide Tests verbessert werden. Die Werte für alle Items im Einzelnen finden sich in Anhang A in den Tabellen 36 und 37.

Tabelle 11. Statistiken der Itemkennwerte vor und nach der Itemselektion der dritten Voruntersuchung zum Subtest TV.

	LESEN 6-7 ($N = 113$)						
	$M_{r_{it}}$	$r_{it_{min}}$	$r_{it_{max}}$	M_p	KR-20	$wMNSQ_{min}$	$wMNSQ_{max}$
vorher	.33	.01	.56	.60	.85	0.83	1.27
nachher	.36	.16	.55	.63	.86	0.86	1.19
	LESEN 8-9 ($N = 189$)						
	$M_{r_{it}}$	$r_{it_{min}}$	$r_{it_{max}}$	M_p	KR-20	$wMNSQ_{min}$	$wMNSQ_{max}$
vorher	.41	.13	.59	.64	.90	0.77	1.33
nachher	.42	.16	.58	.64	.90	0.77	1.26

Durch die Eliminierung weiterer Items wurden beide Tests noch einmal etwas kürzer, was sich vor allem bei LESEN 6-7 mit einer Reduzierung um sechs Items positiv auf die Bearbeitungsdauer und die Zumutbarkeit auswirken sollte. Für die nach der Itemselektion verbleibenden Items wurden über den Item-Fit hinausgehende Voraussetzungen des Rasch-Modells geprüft. Im Einzelnen zählen dazu die Prüfung auf Eindimensionalität, lokale Unabhängigkeit und Personenhomogenität, welche im Folgenden beschrieben werden.

Eindimensionalität. Zur Prüfung der Eindimensionalität wurde eine EFA durchgeführt. Zunächst wurden jedoch die *Voraussetzungen* dafür geprüft. Das Kriterium der Stichprobengröße von mindestens $N = 60$ ist mit $N = 113$ für LESEN 6-7 und $N = 189$ für LESEN 8-9 für beide Tests erfüllt. Zur Schätzung der Itemreliabilität wurde die höchste Korrelation eines Items mit einem anderen Item herangezogen. Bei LESEN 6-7 liegt nur für 16 der Items (also knapp die Hälfte) die höchste Korrelation mit einem anderen Item bei über $r = .60$. Für weitere 13 Items liegt der Wert zwischen $r = .50$ und $r = .60$, für vier Items liegt er zwischen $r = .40$ und $r = .50$, und ein Item liegt bei nur $r = .36$. Der Mittelwert der höchsten Korrelationen liegt bei $r = .59$. Bei

LESEN 8-9 liegt der Mittelwert ebenfalls bei $r = .59$. Auch hier liegt für 18 Items (also knapp die Hälfte) die höchste Korrelation mit einem anderen Item bei über $r = .60$, für 13 Items liegt der Wert zwischen $r = .50$ und $r = .60$, für fünf Items zwischen $r = .40$ und $r = .50$, für zwei Items mit $r = .35$ und $r = .30$ noch darunter. Es handelt sich dabei jedoch um eine Mindestschätzung, und in keinem Test wird der Wert von $r = .20$ unterschritten, was Marcus und Bühner (2009, S. 99) als wichtiges Kriterium anführen.

Der KMO-Wert für LESEN 6-7 kann mit KMO $= .67$ als mäßig gut bezeichnet werden, der Wert von KMO $= .85$ für LESEN 8-9 als gut (vgl. Bühner, 2011, S. 347). Die MSA-Werte belaufen sich bei LESEN 6-7 im Mittel auf einen Wert von MSA $= .65$, was moderat ist, wobei die drei niedrigsten Werte unter MSA $= .50$ liegen und somit eigentlich nicht für eine EFA geeignet sind. Bei LESEN 8-9 liegen alle MSA-Werte über MSA $= .50$, der Mittelwert liegt bei MSA $= .85$. Die MSA-Werte bei LESEN 8-9 sprechen also für eine moderate bis gute Eignung für eine EFA. Der Bartlett-Test wird erwartungsgemäß für beide Tests signifikant, was bedeutet, dass alle Korrelationen signifikant von Null abweichen und die Matrix faktorisierbar ist.

Insgesamt sprechen die meisten Kriterien für eine Eignung der Daten für eine EFA. Lediglich bei den MSA-Werten und der Reliabilitätenschätzung liegen die Werte für einige Items außerhalb des gewünschten Bereichs. Bei der Schätzung der Reliabilität mithilfe der höchsten Korrelation eines Items mit einem anderen Item handelt es sich jedoch nur um eine Mindestschätzung, und auch die MSA-Koeffizienten liefern grundsätzlich lediglich grobe Anhaltspunkte zur Eignung der Items für eine EFA (vgl. Bühner, 2011, S. 344, 348). Da die Stichproben in beiden Fällen über $N = 100$ liegen, sind die Kriterien zudem nicht so streng zu nehmen. Es wurde daher eine EFA durchgeführt, die zumindest einen nützlichen Hinweis darauf geben sollte, ob Eindimensionalität des Subtests TV für beide Tests gegeben ist.

Betrachtet man die *Screeplots*[8] von links nach rechts, ist für beide Tests nach dem ersten Faktor ein deutlicher Knick zu erkennen (s. Abb. 11). Dem Screetest zufolge gibt es also jeweils einen einzigen dominanten Faktor, der zu extrahieren ist.

Die *Parallelanalyse* kommt für LESEN 6-7 ebenfalls auf einen einzigen zu extrahierenden Faktor. Nur der erste Eigenwert der empirischen Eigenwerte liegt über dem entsprechenden Eigenwert am 95 %-Perzentil der simulierten Eigenwerte (s. Abb. 12a), und daher ist nur dieser bedeutsam (vgl. Bühner, 2011). Bei LESEN 8-9 kommt die Parallelanalyse auf zwei zu extrahierende Faktoren (s. Abb. 12b). Bei den Parallelanalysen ist zu berücksichtigen, dass, wie bereits erläutert, andere Berechnungsmethoden verwendet wurden, daher unterscheiden sich auch die Eigenwerte von jenen bei den Screetests.

[8] Bei der Betrachtung der Screeplots fällt auf, dass negative Eigenwerte vorkommen. Dies ist laut B. O. Muthén (http://www.statmodel.com/discussion/messages/8/205.html?1348513453, Zugriff am 13.03.2013) bei kategorialen Variablen und dem WLSMV-Schätzer nicht ungewöhnlich, da mit tetrachorischen Korrelationen gearbeitet wird, die jeweils für einzelne Variablenpaare berechnet werden.

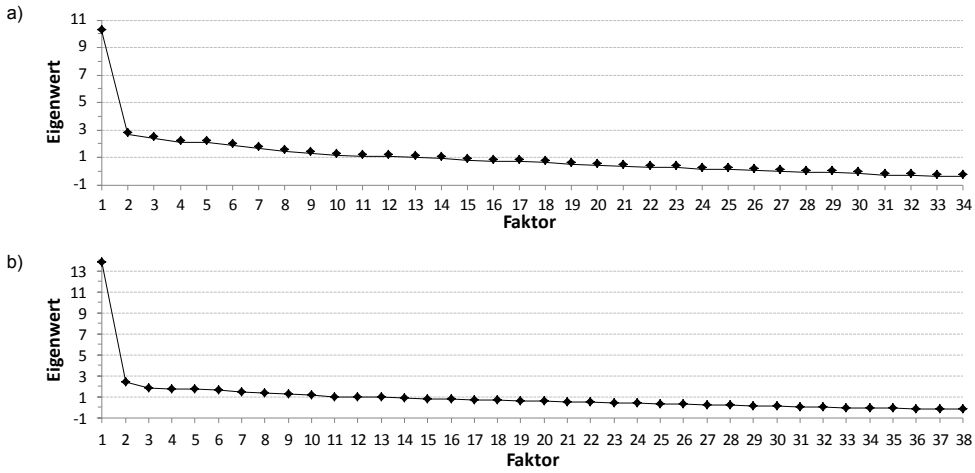


Abbildung 11. Eigenwert-Screeplots für LESEN 6-7 (a) und LESEN 8-9 (b).

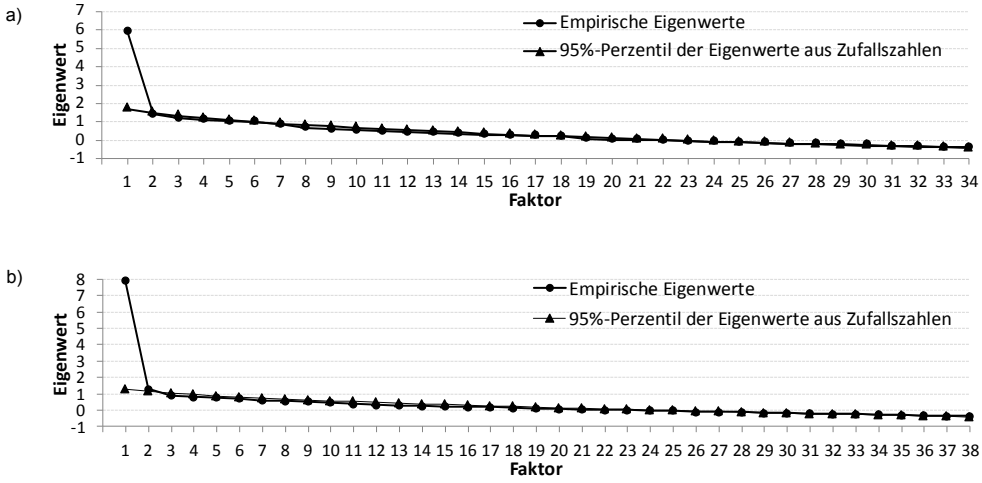


Abbildung 12. Ergebnisse der Parallelanalysen nach Horn für LESEN 6-7 (a) und LESEN 8-9 (b).

Der *MAP-Test* kommt sowohl für LESEN 6-7 als auch für LESEN 8-9 zu dem Ergebnis, dass nur ein einziger Faktor zu extrahieren ist. Bei LESEN 6-7 ergibt sich die kleinste mittlere vierte Potenz der partiellen Korrelation bei einer auspartialisierten Komponente. Der entsprechende Wert liegt bei .0004. Bei LESEN 8-9 fällt die mittlere vierte Potenz der partiellen Korrelation ebenfalls bei einer auspartialisierten Komponente am geringsten aus. Der entsprechende Wert liegt hier bei .0002 (für Details s. Tab. 38 in Anhang A).

Die Betrachtung der Screeplots und die Ergebnisse der MAP-Tests legen für beide Tests eine Ein-Faktorenlösung nahe. Die Parallelanalyse führt bei LESEN 6-7 ebenfalls zu einer Ein-Faktorenlösung, bei LESEN 8-9 jedoch zu einer Zwei-Faktorenlösung. Bei LESEN 6-7 spricht somit alles für eine Ein-Faktorenlösung. Da bei LESEN 8-9 einerseits zwei Kriterien ebenfalls für eine Ein-Faktorenlösung sprechen und andererseits Bühner (2011) bei einem starken ersten Faktor empfiehlt, den MAP-Test gegenüber der Parallelanalyse zu bevorzugen, und die Ein-Faktorenlösung schlussendlich auch leichter zu interpretieren ist als die Zwei-Faktorenlösung, wird diese auch für LESEN 8-9 gewählt.

Nachdem die Anzahl der Faktoren bestimmt war, wurden die *Faktorladungen* betrachtet (s. Tab. 39 in Anhang A). Bei LESEN 6-7 ist die Ladung eines einzigen Items kleiner als $\lambda = .30$. Es handelt sich dabei mit Item 24 um ein Item, das schon zuvor auffällig war und lediglich aufgrund der inhaltlichen Bedeutung für das zu messende Konstrukt nicht eliminiert wurde. Bei LESEN 8-9 ist ebenfalls die Ladung eines einzigen Items kleiner als $\lambda = .30$. Es handelt sich auch hier um das bereits zuvor aufgefallene Item 8.

Insgesamt spricht somit Vieles dafür, dass dem Subtest TV sowohl bei LESEN 6-7 als auch bei LESEN 8-9 erwartungskonform jeweils *eine* latente Variable zugrunde liegt. Somit scheint eine wichtige Voraussetzung für die Rasch-Skalierbarkeit gegeben zu sein.

Lokale Unabhängigkeit. Bei LESEN 6-7 liegen die Q_3-Werte zur Prüfung der lokalen Unabhängigkeit zwischen $Q_3 = .18$ und $Q_3 = .33$, wobei nur drei Werte unterhalb des von Yen (1984) empfohlenen Grenzwertes von $Q_3 = .20$ liegen. Zwar handelt es sich bei dem Grenzwert nur um eine grobe Richtlinie, jedoch bestehen aufgrund dieses Ergebnisses Zweifel am Vorliegen lokaler Unabhängigkeit der Items für diesen Test.

Bei LESEN 8-9 liegen die Q_3-Werte zwischen $Q_3 = .03$ und $Q_3 = .24$, wobei neun Items einen Wert von $Q_3 > .20$ aufweisen. Hier erfüllen zwar deutlich mehr Items das von Yen zur Orientierung angegebene Kriterium, jedoch bei Weitem nicht alle. Allerdings überschreitet kein Item das Kriterium in erheblichem Maße.

Personenhomogenität. Für die *grafischen Modelltests* wurden die Stichproben am Median des Summenwertes in zwei Substichproben aufgeteilt und die Itemparameter für die eine Gruppe an der x-Achse für die andere Gruppe an der y-Achse aufgetragen (s. Abb. 13). Items, die in der Grafik unterhalb der Diagonalen liegen, waren für insgesamt stärkere Schüler mit einem Rohwert oberhalb des Medians (x-Achse) leichter zu lösen als für insgesamt schwächere Schüler mit einem Rohwert unterhalb des Medians (y-Achse). Für Items, die oberhalb der Diagonalen liegen, war das Umgekehrte der Fall. Items, die genau auf der Linie liegen, weisen für beide Subgruppen die gleiche Schwierigkeit auf.

Es wird deutlich, dass sich die Items im Großen und Ganzen entlang der Diagonalen verteilen, jedoch einzelne Items (z. B. bei LESEN 6-7 Item 12 und bei LESEN 8-9 Item 21) relativ weit von der Diagonalen entfernt liegen. Ob diese Abweichungen tatsächlich bedeutsam sind, kann anhand des grafischen Modelltests nicht entschieden werden. Deshalb wurde zusätzlich der Andersen-Test auf Personenhomogenität durchgeführt.

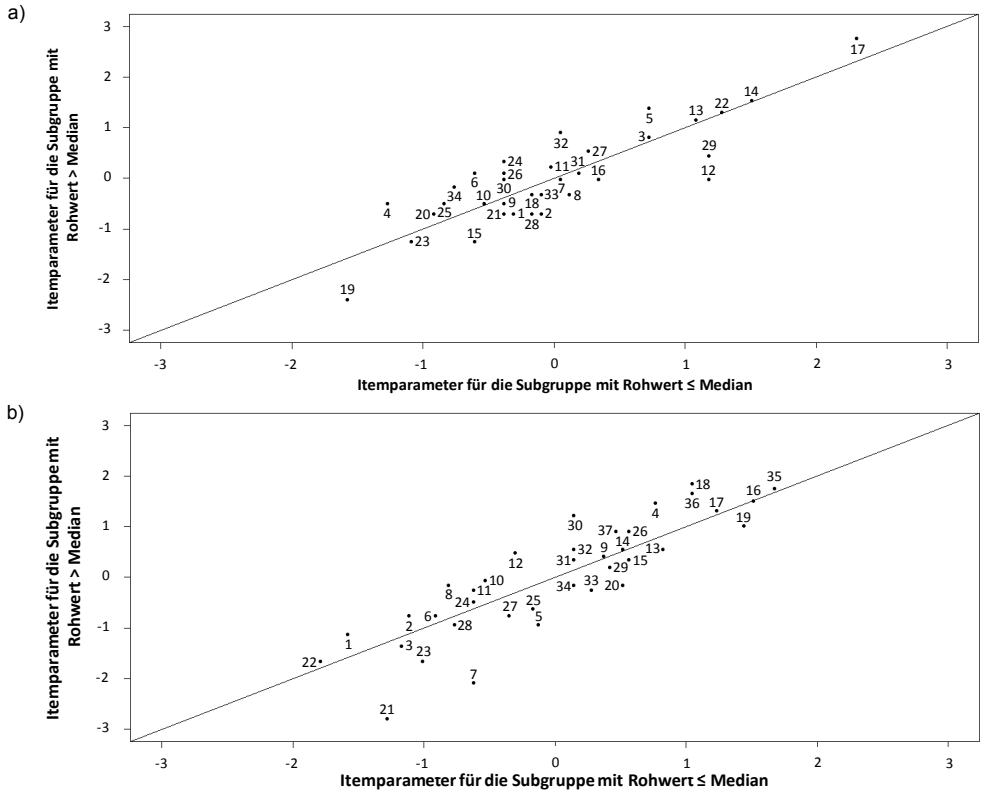


Abbildung 13. Grafischer Modelltest für LESEN 6-7 (a) und LESEN 8-9 (b); Teilungskriterium: Median des Summenwertes; Markierungsbeschriftung: Itemnummer.

Auch beim *Andersen-Test* wurde der Median des Summenwertes als Teilungskriterium gewählt. Das Ergebnis fällt bei LESEN 6-7 nicht signifikant aus ($\chi^2 = 36.21$, $df_{\chi^2} = 33$, $p = 0.32$, $\chi^2_{krit}(p = .80, df = 33) = 39.57$), was für das Vorliegen von Personenhomogenität spricht. Bei LESEN 8-9 dagegen fällt der Andersen-Test signifikant aus ($\chi^2 = 64.40$, $df_{\chi^2} = 37$, $p < .01$, $\chi^2_{krit}(p = .80, df = 37) = 43.98$). Ein signifikantes Ergebnis spricht in diesem Fall gegen das Vorliegen von Personenhomogenität bei LESEN 8-9, was im Hinblick auf die Rasch-Skalierung kritisch zu bewerten ist.

13.4 Beschreibung der Endversionen

Im Folgenden werden die aus den Voruntersuchungen und Revisionen hervorgegangenen Endversionen von LESEN 6-7 und LESEN 8-9 beschrieben.

13.4.1 Testaufbau

LESEN 6-7 und LESEN 8-9 sind analog aufgebaut und bestehen jeweils aus zwei Subtests: einem Subtest zu basaler Lesekompetenz (BLK) und einem Subtest zum Textverständnis (TV). Die Tests werden in Papierform dargeboten, wobei sich beide Subtests in einem gemeinsamen Testheft befinden. Der Subtest BLK besteht aus einer dreiminütigen Satzleseaufgabe und ist für beide Tests identisch. Der Subtest TV besteht für beide Tests aus zwei Texten, jeweils einem expositorischen Text und einem narrativen Text, sowie Verständnisfragen zu den Texten. Die Texte und Fragen unterscheiden sich sowohl inhaltlich als auch in Länge und Schwierigkeit zwischen den Tests. Für die Bearbeitung jedes Textes und der dazugehörigen Fragen stehen in beiden Tests maximal 18 Minuten zur Verfügung.

Insgesamt beträgt die reine Bearbeitungszeit also maximal 39 Minuten: 3 Minuten für den Subtest BLK und 36 Minuten für den Subtest TV, wobei die Dauer für den zweiten Subtest bei schnellen Schülern bzw. schnellen Schulklassen auch kürzer ausfallen kann (s. Kap. 13.4.2). Die Tests sind somit bei zügigem Vorgehen innerhalb einer Schulstunde durchführbar (s. Abb. 14).

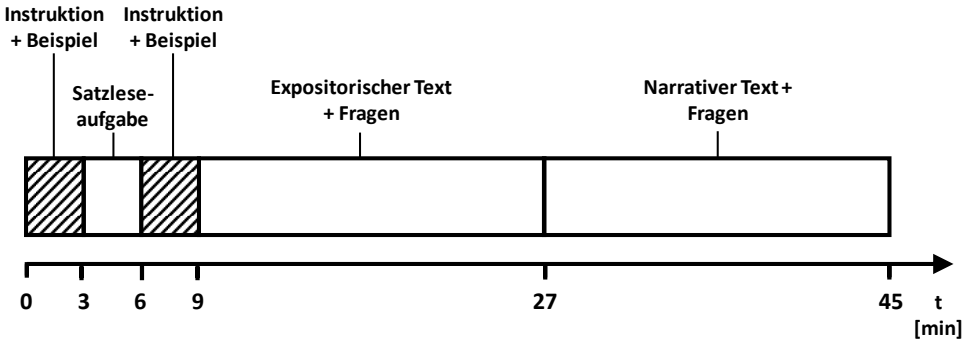


Abbildung 14. Darstellung des Testablaufs (gilt für beide Tests).

Subtest BLK. Der Subtest zur Erfassung der basalen Lesekompetenz ist, wie bereits erwähnt, für beide Tests identisch. Er umfasst eine Liste mit 100 einfachen, kurzen Sätzen, von welchen die Schüler innerhalb von drei Minuten möglichst viele lesen und mit einem Kreuzchen bei „richtig“ oder „falsch“ am Ende der Zeile auf ihre inhaltliche Richtigkeit hin beurteilen sollen. Von den 100 Sätzen sind 54 tatsächlich richtig und 46 tatsächlich falsch. Die Bearbeitungszeit ist so gewählt, dass es kaum ein Schüler

schaffen sollte, alle Sätze zu lesen und korrekt zu beurteilen. Alle Sätze sind nach dem in der Voruntersuchung ermittelten LPQ-Wert absteigend angeordnet, d. h. der leichteste Satz steht an erster, der schwerste an letzter Stelle. Abbildung 15 zeigt ein Beispielitem für diesen Subtest.

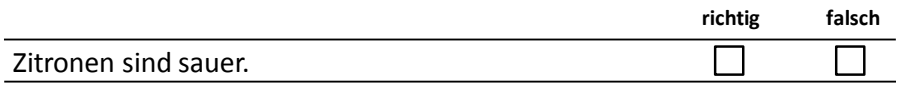


Abbildung 15. Beispielitem für den Subtest BLK.

Subtest TV. Im Subtest zur Erfassung des Textverständnisses werden den Schülern in beiden Tests zwei längere Texte vorgelegt – jeweils ein expositorischer Text und ein narrativer Text. Zu jedem Text werden Verständnisfragen im SC-Format mit jeweils fünf Antwortoptionen gestellt. Die Fragen beziehen sich einerseits auf jeweils eine der zwei Kohärenzebenen (lokal vs. global) sowie andererseits auf eine von zwei Formen der Textrepräsentation (textbasiert vs. Situationsmodell). Darüber hinaus gibt es Fragen, die sich auf die Metaebenen (Textwissen und Darstellungsstrategien) sowie kontextualisiertes Wortverständnis beziehen.

Bei LESEN 6-7 umfasst der expositorische Text 555 Wörter, der narrative Text 705 Wörter. Zu jedem Text werden 17 Verständnisfragen gestellt. Somit enthält der gesamte Subtest 34 Items (Fragen). Die Beantwortung von 13 Fragen erfordert die Bildung lokaler Kohärenzen und die Beantwortung von 15 Fragen die Bildung globaler Kohärenzen. Die Antworten auf 9 Fragen sind textbasiert und 19 Fragen erfordern den Aufbau eines Situationsmodells. Darüber hinaus prüfen 5 Fragen kontextualisiertes Wortverständnis und 1 Frage prüft die Metaebenen (s. Tab. 12).

Bei LESEN 8-9 umfasst der expositorische Text 706 Wörter, der narrative Text 897 Wörter. Beide Texte sind somit etwas länger als die Texte von LESEN 6-7. Zu jedem Text werden bei LESEN 8-9 19 Verständnisfragen gestellt. Insgesamt enthält der Subtest TV also 38 Items (Fragen). Bei LESEN 8-9 erfordern 21 Fragen die Bildung lokaler Kohärenz und 7 Fragen die Bildung globaler Kohärenz. Die Antworten auf 5 Fragen sind textbasiert, während die Antworten auf 23 Fragen den Aufbau eines Situationsmodells erfordern. Darüber hinaus prüfen 6 Fragen kontextualisiertes Wortverständnis sowie 4 Fragen die Metaebenen (s. Tab. 12).

Die Betrachtung der Itemverteilung über die verschiedenen Kohärenzebenen und Repräsentationsformen zeigt, dass alle anvisierten Kohärenzebenen und Repräsentationsformen in beiden Tests vorhanden sind (rein textbasierte globale Kohärenzbildung zu testen, erschien von Anfang an nicht möglich). Zugleich wird aber auch deutlich, dass die angestrebte gleichmäßige Verteilung der Items über die Kohärenzebenen und Repräsentationsformen bei beiden Tests nicht erreicht wurde. Tabelle 40 in Anhang A gibt für jedes Item der Endversionen beider Tests an, auf welche Verständnisebene es sich bezieht.

Tabelle 12. Verteilung der Items des Subtests TV über die Verständnisebenen und Repräsentationsformen.

LESEN 6-7			
	Lokale Kohärenz	Globale Kohärenz	Σ
Textbasiert	9	–	9
Situationsmodell	4	15	19
Σ	13	15	28
Wortverständnis im Kontext		5	
Metaebenen		1	
Gesamt:		34	

LESEN 8-9			
	Lokale Kohärenz	Globale Kohärenz	Σ
Textbasiert	5	–	5
Situationsmodell	16	7	23
Σ	21	7	28
Wortverständnis im Kontext		6	
Metaebenen		4	
Gesamt:		38	

13.4.2 Durchführung

Im Folgenden wird grob beschrieben, wie LESEN 6-7 und LESEN 8-9 durchzuführen sind, und wie sie auch im Rahmen der Normierung, der Validierung und der weiteren empirischen Erprobung durchgeführt wurden. Eine detailliertere Beschreibung findet sich in den Testmanualen (Bäuerlein, Lenhard & Schneider, 2012a, 2012b). Einige Aspekte (z. B. Zielgruppe, Durchführungsdauer) wurden bereits angesprochen. Sie werden daher an dieser Stelle nicht noch einmal wiederholt. Im Folgenden wird nur noch auf das Testmaterial, die Durchführungsbedingungen und die Instruktionen eingegangen.

Während der Testung sollten sich nur die benötigten Testmaterialien auf den Arbeitsplätzen der Schüler befinden. Zudem ist für Ruhe und angemessene Lichtverhältnisse zu sorgen. Folgende Materialien sind für die Testdurchführung nötig: Ein Testbogen pro Schüler, zwei Stifte pro Schüler (einer davon als Ersatzstift), ein Testbogen für Demonstrationszwecke für den Testleiter, die Testinstruktion und ein Zeitmessgerät (z. B. Stoppuhr). Weitere Hilfsmittel sind nicht gestattet. Da die Ergebnisse von LESEN 6-7 und LESEN 8-9 nur dann aussagekräftig sind, wenn die Testungen unter standardisierten Bedingungen durchgeführt werden, ist die im Manual vorgegebene Instruktion zu verwenden. Weiter ist es wichtig, die maximale Bearbeitungsdauer für jeden Testabschnitt genau einzuhalten. Beim Subtest BLK ist keine Verkürzung

möglich. Falls beim Subtest TV alle Schüler mit der Bearbeitung der Fragen zu einem Text bereits vorzeitig fertig sind, darf früher zum nächsten Text weitergegangen bzw. darf dann beim zweiten Text der Test früher beendet werden. Es dürfen keine über die Instruktion hinausgehenden inhaltlichen Hilfestellungen gegeben werden. Abschreiben und jegliche Art der Kommunikation der Schüler untereinander sind zu verhindern. Die Schüler dürfen auch nur am jeweils aktuellen Subtest arbeiten, ein Vor- oder Zurückblättern zu anderen Subtests ist nicht gestattet.

Die Instruktionen werden im Folgenden am Beispiel von LESEN 8-9 dargestellt (für LESEN 6-7 unterscheiden sich lediglich die Seitenangaben). Die kursiv gedruckten Textstellen sind wortwörtlich wiederzugeben, bei den übrigen Abschnitten handelt es sich um Hinweise für den Testleiter. Im Fall einer Einzeltestung sind die Instruktionen entsprechend anzupassen.

1. Allgemeine Instruktion:
Ich werde heute mit euch eine Leseuntersuchung durchführen. Es ist dabei wichtig, dass ihr euch anstrengt und bei allen Aufgaben euer Bestes gebt! Ich werde euch nun Hefte mit Leseaufgaben austeilen. Lasst diese bitte noch geschlossen und wartet, bis alle ein Heft bekommen haben. Wir werden dann gemeinsam die erste Seite ausfüllen.

2. Personenbezogene Daten:
Auf der ersten Seite sind personenbezogene Angaben einzutragen. Hierzu gehören Vor- und Nachname, Schule und Klasse, Geburtsdatum, Geschlecht, das aktuelle Datum und die Muttersprache. Wenn der Schüler zu Hause mit einem oder beiden Elternteilen eine oder mehrere andere Sprachen als Deutsch spricht, sind diese hier zu vermerken, andernfalls wird „Deutsch“ eingetragen. Außerdem ist das entsprechende Kästchen anzukreuzen, falls bei dem Schüler eine LRS-Diagnose vorliegt. Ansonsten ist das Kästchen freizulassen.

3. Instruktion Basale Lesekompetenz:
Jetzt blättert bitte alle einmal um, sodass ihr auf Seite 3 seid. Hier wird zunächst erklärt, was ihr machen sollt. Ich lese es euch einmal vor, passt bitte gut auf!
Bitte die Instruktion im Testheft vorlesen und anschließend gemeinsam die Beispielsätze bearbeiten. Dabei darauf achten, dass noch keiner weiterblättert.
Auf den nächsten Seiten kommen ganz viele solcher Sätze und ihr sollt versuchen, innerhalb von drei Minuten so viele wie möglich zu bearbeiten. Es sind so viele Sätze, dass es wahrscheinlich keiner von euch schaffen wird, alle zu lesen. Das macht überhaupt nichts! Wichtig ist, dass jeder sich anstrengt und so viele bearbeitet, wie er kann. Bitte denkt daran, euch nicht zu korrigieren! Wenn ich „los“ sage, fangt ihr an und wenn ich sage „stopp“, hört ihr sofort auf zu schreiben und legt die Stifte weg. Hat noch jemand eine Frage dazu?
Wenn es keine Fragen mehr gibt:
Los!

Nach 3 Minuten:
Stopp! Legt bitte alle die Stifte weg und blättert auf Seite 9.

4. Instruktion Textverständnis:
Wir machen gleich weiter mit der nächsten Aufgabe. Ich lese euch wieder vor, was ihr machen sollt. Lest bitte mit!
Bitte die Instruktion, den Beispieltext sowie die Frage und die Antwortalternativen im Testheft vorlesen und die Schüler fragen, welche Antwort die richtige ist. Bei der richtigen Antwort:
Genau, die letzte Antwort ist richtig! Denn im Text steht „Damit sind sie die schwersten Tiere, die jemals auf der Erde gelebt haben." D. h. die Blauwale sind auch schwerer als die Dinosaurier waren. Diese Beispielaufgabe soll euch zeigen, dass die Antworten nicht immer wörtlich im Text stehen, sondern dass ihr auch manchmal ein bisschen überlegen müsst. Bitte bearbeitet zunächst nur den ersten Text und die Fragen dazu. Es sind dafür 18 Minuten Zeit. Ihr dürft jederzeit noch einmal im Text nachlesen. Ich sage euch Bescheid, wenn ihr zum nächsten Text übergehen sollt. Hat jemand noch eine Frage dazu?
Wenn es keine Fragen mehr gibt:
Bitte blättert jetzt um und bearbeitet den ersten Text und die Fragen dazu.
Nach 18 Minuten:
Blättert jetzt bitte alle zur Seite 16 und bearbeitet den zweiten Text. Ihr habt dafür wieder 18 Minuten Zeit.
Nach 18 Minuten:
Die Zeit ist um, die Leseuntersuchung ist zu Ende. Bitte legt eure Stifte weg und schließt die Hefte. Ich werde sie gleich einsammeln. Vielen Dank, dass ihr so gut mitgemacht habt!

Hält sich der Testleiter an den Instruktionstext, werden den Testpersonen vor den Subtests Itembeispiele vorgegeben, außerdem werden die Testpersonen über die Bearbeitungsdauer informiert, was u. a. von J. Rost (2004, S. 76) empfohlen wird. Darüber hinaus gibt das Testmanual Hinweise, wie der Testleiter sich bei Fragen der Schüler während der Testung verhalten soll:

Wenn im Laufe der Instruktion oder im Anschluss an einen Subtest Fragen auftreten, sind diese durch den Testleiter zu beantworten, sofern diese sich auf die Form der Aufgabenbeantwortung und nicht auf den Inhalt beziehen. Treten während der Testung Fragen auf, ist im Falle einer Gruppentestung eine Störung der anderen Schüler zu vermeiden. Stattdessen sollte in diesem Fall die Frage leise mit dem Schüler geklärt werden, sofern diese für die Testdurchführung relevant ist. Auf inhaltliche Fragen kann Folgendes geantwortet werden:
Schau dir die Frage noch einmal genau an und schau noch einmal im Text nach, ob du die Antwort findest oder sie dir erschließen kannst. Wenn du die Lösung dann immer noch nicht weißt, wähle die Antwort, von der du denkst, dass sie am ehesten stimmen könnte oder überspringe die Aufgabe und versuche am Ende noch einmal sie zu lösen.

13.4.3 Auswertung

Unabhängig von der Fragestellung, welche mit dem Einsatz der Tests beantwortet werden soll, ist zunächst für beide Subtests ein Rohwert zu bestimmen. Sowohl bei LESEN 6-7 als auch bei LESEN 8-9 wird bei beiden Subtests für jedes korrekt gelöste Item 1 Punkt vergeben, für jeden Fehler und jedes nicht bearbeitete Item 0 Punkte. Durch Aufsummieren der Punkte innerhalb der Subtests erhält man die Subtest-Rohwerte. Um eine ökonomische Auswertung zu gewährleisten, Auswertungsfehler zu vermeiden und die Auswertungsobjektivität sicherzustellen, stehen für die Ermittlung der Rohwerte Schablonen bzw. Auswertungsfolien zur Verfügung. Aufgrund des geschlossenen Antwortformats ist prinzipiell auch eine Auswertung anhand eines Dokumentenscanners möglich. Die Rohwerte können auf einem übersichtlichen Auswertungsbogen eingetragen werden.

Für die einzelnen Klassenstufen sowie nach Klassenstufe und Schulart (Hauptschule, Realschule und Gymnasium) getrennt stehen in den Testmanualen Normtabellen zur Verfügung. Diese basieren auf den Daten einer umfangreichen Normierung, die in Kapitel 14 noch ausführlich beschrieben wird. In den Tabellen ist jedem Rohwert ein Prozentrang, ein T-Wert und ein T-Wertband zugeordnet. Nach der Ermittlung der Normwerte für beide Subtests, kann zusätzlich ein Gesamtwert gebildet werden, der sich aus der Summe der beiden Subtest-T-Werte ergibt. Durch die Verwendung der T-Werte (anstatt der Rohwerte) der Subtests bei der Summenbildung werden beide Subtests gleich gewichtet. Die Summe der Subtest-T-Werte stellt somit den Rohwert des Gesamtergebnisses dar. Für dieses werden dann wiederum Prozentränge, T-Werte und T-Wertbänder angegeben.

Darüber hinaus kann ein Profildiagramm erstellt werden, in das die Subtestergebnisse und das Gesamtergebnis eingetragen werden können, um Stärken und Schwächen einzelner Schüler besser erkennen zu können. Schließlich werden Vergleichswerte für Subtestdifferenzen zur Verfügung gestellt, mit deren Hilfe geprüft werden kann, ob sich die T-Werte zwischen den beiden Subtests nur in gewöhnlichem Maß oder aber außergewöhnlich stark unterscheiden. Hierauf wird in Kapitel 14.4 noch genauer eingegangen.

13.5 Diskussion

Im vergangenen Kapitel wurde die Konstruktion von LESEN 6-7 und LESEN 8-9 in allen einzelnen Schritten dargestellt. Im Folgenden werden das methodische Vorgehen und die Ergebnisse zunächst zusammengefasst und anschließend kritisch diskutiert.

Zusammenfassung. Die Entwürfe für LESEN 6-7 und LESEN 8-9 resultierten aus Implikationen aus Theorien und Befunden zum Leseverständnis in der Sekundarstufe sowie aus testtheoretischen Überlegungen. Für die beiden analog aufgebauten Tests

wurden je zwei Subtests entworfen. Für den Subtest BLK, der für beide Tests identisch ist, wurde eine Satzleseaufgabe ausgewählt. Für den Subtest TV wurden für jeden Test je ein expositorischer Text und je ein narrativer Text ausgewählt und jeweils zahlreiche Verständnisfragen im SC-Format dazu formuliert, die sich auf verschiedene Ebenen des Textverständnisses bezogen. Die Testentwürfe wurden in mehreren Voruntersuchungen empirisch erprobt und jeweils anschließend revidiert.

Die Voruntersuchung für den Subtest BLK fand PC-gestützt statt. Die Itemselektion basierte auf LPQ-Werten, die die Korrektheit der Lösung sowie die Bearbeitungsgeschwindigkeit für jedes Item berücksichtigten. Aus 140 Items wurden anhand der LPQ-Werte diejenigen 100 Items ausgewählt, welche am schnellsten korrekt beantwortet wurden. Die Ergebnisse der Itemanalyse sprechen dafür, dass die ausgewählten Items ausreichend leicht sind, um zwischen den Schülern hauptsächlich auf der Basis der Lesegeschwindigkeit zu differenzieren (und nicht z. B. aufgrund von Vorwissen).

Für den Subtest TV erfolgte die Itemselektion in der ersten Voruntersuchung auf Basis von KTT-Kennwerten (Itemschwierigkeit, Trennschärfe, Selektionskennwert), inhaltlichen Überlegungen sowie aufgrund von Experten- und Zielgruppenbefragungen. Zudem wurde die interne Konsistenz betrachtet. In der zweiten Voruntersuchung wurde geprüft, ob die Kombination von zwei Texten und Verständnisfragen eine angemessene und zumutbare Aufgabenstellung für die Zielgruppe darstellt. In der dritten Voruntersuchung wurden über die KTT-Kennwerte hinaus auch Rasch-Kennwerte (wMNSQ) für die Itemselektion herangezogen. Im Verlauf der Voruntersuchungen zum Subtest TV wurden zum einen die Texte angepasst und zum anderen die Fragen und Distraktoren optimiert, selektiert und ausgetauscht, bis schließlich weitgehend zufriedenstellende Itemkennwerte und Reliabilitätswerte erreicht waren. Zudem wurde in der dritten Voruntersuchung die strukturelle Schwierigkeit der Texte in ihrer Endversion anhand von Lesbarkeitsindizes beurteilt sowie nach der Itemselektion verschiedene über den Item-Fit hinaus gehende Voraussetzungen für die Rasch-Skalierung (Eindimensionalität, lokale Unabhängigkeit, Personenhomogenität) geprüft.

Die Untersuchung der strukturellen Schwierigkeit der Texte zeigte, dass die expositorischen Texte für die Zielgruppen eher schwer zu verstehen sein dürften, die narrativen Texte dagegen eher leicht. Das bedeutet, dass in jedem Test jeweils ein strukturell anspruchsvoller und ein strukturell eher einfacher Text enthalten ist, sodass jeweils ein breites Schwierigkeitsspektrum abgedeckt ist.

Nach der Itemselektion im Rahmen der dritten Voruntersuchung verblieben 34 Items (17 Items pro Text) für LESEN 6-7 und 38 Items (19 Items pro Text) für LESEN 8-9. Die Rohwertverteilungen fielen für beide Tests zufriedenstellend aus. Es wurden weder Boden- noch ausgeprägte Deckeneffekte festgestellt – wenn auch die Verteilungen in den Gymnasialklassen leicht linksschief ausfielen. Die KTT-Itemkennwerte fielen nach der dritten Voruntersuchung ebenfalls größtenteils zufriedenstellend aus. Bezüglich des Selektionskennwertes weist bei jedem Test jeweils ein Item einen niedrigen Wert

von SK $< .20$ auf, alle anderen SK-Werte liegen wie eigentlich erwünscht darüber. Die Schwierigkeitsindizes liegen bei LESEN 6-7 zwischen $p = .17$ und $p = .89$, bei LESEN 8-9 zwischen $p = .33$ und $p = .89$. Somit decken die Items jeweils wie erwünscht ein breites Schwierigkeitsspektrum ab. Die interne Konsistenz ist bei LESEN 6-7 mit KR-20 $= .86$ als gut, bei LESEN 8-9 mit KR-20 $= .90$ als sehr gut zu bezeichnen.

Im Hinblick auf die Rasch-Analysen fallen die Ergebnisse inkonsistenter aus. Für beide Tests liegen die wMNSQ-Werte im erwünschten Bereich zwischen 0.75 und 1.33. Jedoch liegt bei LESEN 6-7 für zwei wMNSQ-Werte der *T*-Wert bei 2, und bei LESEN 8-9 liegen zwei *T*-Werte sogar darüber. Diese Werte sind kritisch und weisen auf signifikante Abweichungen der entsprechenden Items vom Modell hin. Bezüglich der Dimensionalität spricht bei beiden Tests Vieles für das Vorliegen von Eindimensionalität. Darüber hinaus weisen die Q_3-Werte bei LESEN 6-7 darauf hin, dass die lokale Unabhängigkeit der Items verletzt ist, während bei LESEN 8-9 die entsprechenden Q_3-Werte grenzwertig ausfallen. Das Ergebnis des Andersen-Tests spricht bei LESEN 6-7 für ein Vorliegen von Personenhomogenität, bei LESEN 8-9 jedoch dagegen.

Kritische Betrachtung. Im Folgenden sollen zunächst einige Kritikpunkte im Hinblick auf den Testaufbau und die Testinhalte diskutiert werden, bevor auf die Voruntersuchungen eingegangen wird.

In Bezug auf die Inhalte von LESEN 6-7 und LESEN 8-9 kann kritisch angemerkt werden, dass die Lesekompetenz nicht umfassend abgebildet wird. Affektive, motivationale und soziale Aspekte des Lesens werden nicht explizit berücksichtigt. Die Erfassung von über die kognitiven Prozesse hinausgehenden Aspekten des Lesens stand im Widerspruch zu dem Ziel, eine Schulstunde nicht zu überschreiten und einen ökonomischen Test zu entwickeln, der gut in den Schulalltag integrierbar ist. Zudem würde ein längerer Test die Schüler möglicherweise überfordern und die Zumutbarkeit beeinträchtigen. Daher fokussieren die Tests auf die kognitiven Aspekte des Lesens, die den Kern der Lesekompetenz sowie der Literalität darstellen. In Bezug auf die Erfassung der noch umfassenderen Literalität erscheint es zudem fraglich, ob es überhaupt möglich wäre, im Rahmen einer Testung authentische Leseziele und -motivation zu induzieren und somit z. B. die Fähigkeit zur Nutzung von Texten für eigene Ziele zu prüfen. Auf das „Lesen“ von Grafiken oder Diagrammmen wurde ebenfalls verzichtet, da möglichst reines Leseverständnis erfasst werden sollte und nicht zusätzlich z. B. Bildverständnis oder mathematisches Verständnis. Zudem zeigte sich in der DESI-Studie, dass im Unterricht Tabellen und Schaubilder – insbesondere an Gymnasien – eine deutlich geringere Rolle spielen. Vielmehr werden vor allem Sachtexte und literarische Texte verwendet (DESI-Konsortium, 2006, S. 32).

Für den Subtest BLK von LESEN 6-7 und LESEN 8-9 wurde eine stille Satzleseaufgabe gewählt. Zwar handelt es sich bei Stillleseaufgaben um die ökologisch validere Variante im Vergleich zu Lautleseaufgaben, und sie ermöglichen den Einsatz der Tests als Gruppentests, jedoch ist es beim stillen Lesen nicht möglich, die basale Lesekompetenz

ohne Verständnisüberprüfung zu erfassen. Die Beurteilung der Sätze hinsichtlich ihrer inhaltlichen Richtigkeit im Subtest BLK ist nötig, um sicherzugehen, dass die Schüler die Sätze auch tatsächlich lesen. Andere Gruppentests, die basale Lesekompetenz oder Lesegeschwindigkeit anhand einer stillen Leseaufgabe erfassen (z. B. SLS 5-8), weisen die gleiche Problematik auf. Das Verständnis einfacher kurzer Sätze ist jedoch noch nicht als hierarchiehöhere Verständnisleistung anzusehen, weshalb diese Art der Operationalisierung vertretbar ist.

Darüber hinaus sind bei den Testentwürfen zum Subtest TV auch die ausgewählten Verständnisfragen kritisch zu betrachten. So hätten noch gezielter verschiedene Ebenen des Situationsmodells (z. B. Ziel, Emotion, Protagonist) und Inferenzarten berücksichtigt werden können, um hier ein wirklich umfassendes Bild der Verständnisleistung zu bekommen. Dies hätte jedoch noch eine deutlich höhere Itemanzahl erforderlich gemacht und die Testzeit verlängert. Zudem wäre es noch schwieriger gewesen bei der Itemselektion darauf zu achten, dass alle verschiedenen Dimensionen von Situationsmodellen noch in der Endversion vertreten sind.

Einen weiteren Kompromisspunkt in Bezug auf die Items stellt das gebundene Antwortformat dar. Es bringt zwar den Vorteil mit sich, dass die Testdurchführung und -auswertung ökonomisch und objektiv ablaufen kann und dass nicht zusätzlich Schreibkompetenzen überprüft werden. Zugleich birgt es jedoch den Nachteil, dass nur ein Wiedererkennen erforderlich ist und keine freie Reproduktion. Problematisch dabei ist auch, dass die Antwortalternativen bereits Hinweise auf die richtige Lösung enthalten können und qualitativ unterschiedliche Lösungsstrategien möglich sind, die die Dimensionalität des Tests beeinflussen können, d. h. es werden eventuell mehrere Fähigkeiten gemessen und nicht nur eine (Bühner, 2011). Ein Hinweis in den Instruktionen, dass die Schüler keine Frage auslassen und bei Nichtwissen raten sollen, hätte dazu beitragen können, dass bei Nichtwissen alle Schüler die gleiche Strategie anwenden. Das Fehlen eines Hinweises dazu, wie Schüler sich verhalten sollen, wenn sie eine Frage nicht beantworten können, kann dazu führen, dass manche Schüler raten, während andere die Frage auslassen. Ein MC-Antwortformat, bei dem die Anzahl der richtigen Antworten unbekannt ist, hätte zudem im Gegensatz zum gewählten SC-Format verhindern können, dass zusätzlich zur direkten Wahl der korrekten Antwort eine Beantwortung der Fragen nach dem Ausschlussverfahren möglich ist. Außerdem hätte durch MC-Fragen die Ratewahrscheinlichkeit deutlich gesenkt werden können. Allerdings dürfte Schülern das SC-Format am vertrautesten sein. Sie kennen dieses z. B. aus Quizsendungen im Fernsehen. Durch die Verwendung eines vertrauten Antwortformats sollten Fehler aufgrund falsch verstandener Instruktionen vermieden werden. Generell besteht jedoch bei gebundenen Antwortformaten das Problem, dass bei ihnen stets eine gewisse Lösungswahrscheinlichkeit aufgrund von Raten gegeben ist. Die Höhe dieser Wahrscheinlichkeit hängt dabei von der Anzahl der Antwortalternativen und der Qualität der Distraktoren ab (vgl. Bühner, 2011), worauf bei der Testkonstruktion der Subtests TV großer Wert gelegt wurde. Es wurde eine große Anzahl

an Antwortalternativen gewählt, und die Distraktoren wurden mehrfach überarbeitet, um eine hohe Qualität zu erreichen. Auch wenn aufgrund des geschlossenen Antwortformats und vor allem beim Subtest BLK aufgrund des richtig-falsch-Formats (also zwei Antwortoptionen) eine gewisse Lösungswahrscheinlichkeit aufgrund von Raten gegeben ist, wurde darauf verzichtet, falsche Lösungen mit Minuspunkten zu „bestrafen“, da sich dies sehr verunsichernd auf die Schüler auswirken kann (vgl. Bühner, 2011, S. 122).

Die gewählte Variante, beim Subtest TV den Schülern bei der Bearbeitung der Fragen das Zurückblättern zum Text zu erlauben, wird nicht von allen Autoren als die ökologisch validere angesehen, da in der Realität die Texte selten mehrmals gelesen würden. Für LESEN 6-7 und LESEN 8-9 erschien es jedoch wichtiger, die Konfundierung mit der Gedächtnisleistung zu vermeiden bzw. zu reduzieren. Zudem ist in der Realität – auch wenn sie selten genutzt werden mag – die Möglichkeit des Zurückblätterns meist gegeben.

Weiterhin ist in Bezug auf die Textauswahl für den Subtest TV anzumerken, dass die expositorischen Texte laut Lesbarkeitsindizes für die Zielgruppen als strukturell relativ schwierig einzustufen sind und die narrativen Texte sind aufgrund der Berücksichtigung des Urheberrechts sehr alt. Somit entsprechen die gewählten Texte nicht unbedingt authentischen Freizeit-Lesesituationen von Schülern der Zielgruppe. In Schullesebüchern findet man aber in den Klassenstufen sechs bis neun durchaus ähnliche Texte (z. B. auch Texte von Johann Peter Habel).

Darüber hinaus sind hinsichtlich der Voruntersuchungen einige Punkte kritisch zu diskutieren. Generell ist diesbezüglich zu bemängeln, dass alle Voruntersuchungen ausschließlich im Regierungsbezirk Unterfranken stattfanden. Bei einem Test, der für alle deutschen Bundesländer geeignet sein soll, könnte dies problematisch sein, da bekannt ist, dass sich die Leseleistung der Sekundarschüler zwischen den Bundesländern deutlich unterscheidet (vgl. Kap. 6). Aus praktisch-organisatorischen und ökonomischen Gründen konnten jedoch zunächst nicht mehr Regionen in die Untersuchungen einbezogen werden.

Im Hinblick auf die Voruntersuchung zum Subtest BLK ist zudem zu bemängeln, dass nur eine einzige Voruntersuchung durchgeführt wurde, und diese auch nur an einer sehr kleinen Stichprobe, die nicht das ganze Leistungsspektrum der eigentlichen Zielgruppen umfasste. Hinzu kommt, dass der Vortest computerbasiert stattfand, sodass die Ergebnisse möglicherweise nicht eins zu eins auf die Endversion der Tests in Papierform übertragbar sind. Außerdem ist kritisch anzumerken, dass die ausgewählten Sätze den Zielgruppen nie ohne Zeitbegrenzung vorgegeben wurden, sodass nicht empirisch belegt ist, dass bei unbegrenzter Bearbeitungszeit alle Sätze tatsächlich von allen Schülern korrekt gelöst werden.

Für den Subtest TV wurden ebenfalls in keiner der drei Voruntersuchungen alle Subgruppen (Schularten und Klassenstufen) der Zielgruppen eingeschlossen. Weiter erfolgte beim Subtest TV die Itemselektion nicht ausschließlich aufgrund statistischer

Kennwerte, sondern auch aufgrund inhaltlicher Überlegungen. Einige Items, die die KTT- und IRT-Kriterien nicht erfüllten, wurden aufgrund ihrer inhaltlichen Bedeutung im Test behalten, um das Zielkonstrukt angemessen abzubilden und inhaltliche Validität der Tests zu gewährleisten. Als inhaltliches Kriterium zählte z. B., dass die verschiedenen Verständnisebenen vertreten sein sollten. Durch Umformulierungen und Änderungen der Distraktoren wurde jedoch versucht, die statistischen Kennwerte der Items zu verbessern, ohne sie inhaltlich zu verändern. Allerdings gelang es dennoch nicht, die verschiedenen Verständnisebenen gleichmäßig über die Items zu verteilen, wodurch die Konstruktvalidität möglicherweise doch beeinträchtigt ist – insbesondere bei LESEN 6-7, da dort beispielsweise nur ein Item zur Prüfung der Metaebenen verblieben ist. Darüber hinaus ist es problematisch, dass die Zuordnung der Items des Subtests TV zu den Verständnisebenen nicht immer eindeutig getroffen werden konnte. Beispielsweise kann eine Beantwortung von Fragen mithilfe von Vorwissen nicht immer ausgeschlossen werden.

Bei den Itemrevisionen für den Subtest TV wurden auch Anmerkungen von Experten sowie der Zielgruppen berücksichtigt. Hierbei wäre es sinnvoll gewesen, systematische kognitive Interviews zur Itemüberprüfung durchzuführen (vgl. z. B. Prüfer & Rexroth, 2005), anstatt unsystematisch und ohne Dokumentation die Experten und die Zielgruppe zu den Texten und Items zu befragen. Für die Zielgruppe wäre darüber hinaus eine Durchführung des Tests mit der sogenannten „Think-aloud-Technik“ aufschlussreich gewesen. Aufgrund eines entsprechend erhöhten Zeitaufwands wurde jedoch darauf verzichtet.

Trotz aller Kritikpunkte an der Testkonstruktion sind die Ergebnisse nach der dritten Voruntersuchung im Großen und Ganzen zufriedenstellend, wenngleich einzelne Items des Subtests TV auch am Ende noch nicht modellkonform zu sein scheinen. Diese Items wurden jedoch aufgrund ihrer inhaltlichen Bedeutsamkeit nicht eliminiert. Eine erneute Überprüfung der Itemgüte und der Modellpassung auf Basis der im Rahmen der im folgenden Kapitel beschriebenen Normierung erhobenen Daten wird zeigen, ob sich die kritischen Items bei einer größeren Stichprobe, die das gesamte Leistungsspektrum der Zielpopulationen abdeckt, noch bewähren. Alle nach der Itemselektion berechneten Kennwerte sind ohnehin an einer neuen Stichprobe zu bestätigen.

Kapitel 14

Normierung

Die Normierung von LESEN 6-7 und LESEN 8-9 hatte zum Ziel, Vergleichswerte zur Verfügung zu stellen, mit denen künftig die Ergebnisse von Schülern verglichen werden können (vgl. Kap. 13.4.3). Die hierfür herangezogenen Normstichproben sollten ausreichend groß und möglichst repräsentativ für die Zielpopulationen (Klassenstufen sechs und sieben bzw. acht und neun der deutschen Regelschulen) sein. Da Vorbefunde darauf hinwiesen, dass sowohl zwischen den Bundesländern als auch zwischen den Schularten deutliche Unterschiede in der Leseleistung bestehen (vgl. Kap. 6), sollten möglichst große Stichproben aus möglichst vielen Bundesländern und den wichtigsten Schularten (Hauptschule, Realschule und Gymnasium) rekrutiert werden.

Im Folgenden werden zunächst die Datenerhebungen beschrieben. Anschließend wird die Zusammensetzung der Normstichproben dargestellt, um daraufhin deren Größe und Repräsentativität kritisch zu prüfen. Es folgen deskriptive Statistiken und eine Betrachtung der Rohwertverteilungen, bevor auf die Berechnung der Normwerte und die Erstellung der Normtabellen eingegangen wird.

14.1 Datenerhebung

Das Ziehen einer Zufallsstichprobe aus der Gesamtpopulation aller Sechst- bis Neuntklässler deutscher Regelschulen oder selbst einer Klumpenstichprobe war schon aufgrund der Freiwilligkeit der Teilnahme und der begrenzten finanziellen und zeitlichen Ressourcen des Projekts nicht möglich. Mit der Unterstützung einiger Mitglieder des Hogrefe Schultest-Herausgebergremiums konnten jedoch in mehreren deutschen Bundesländern Schulen für die Teilnahme an der Normierung rekrutiert werden. Es wurde ein Plan erstellt, wie viele Schulklassen welcher Klassenstufen und Schularten an den verschiedenen Standorten getestet werden sollen. Dabei wurde darauf geachtet, dass alle Schularten und Klassenstufen möglichst gleichmäßig über Deutschland verteilt waren. Leider ließ sich dieser Plan nicht einhalten, da einige Bundesländer die Datenerhebung gar nicht oder nur in bestimmten von ihnen vorgegebenen Klassenstufen und Schularten genehmigten. Daher handelt es sich letztendlich um Verfügbarkeitsstichproben.

Als Normierungszeitpunkt wurde das Schuljahresende gewählt. Die Datenerhebung fand entsprechend in den letzten beiden Monaten des Schuljahres 2009/2010 statt.

Abhängig vom Bundesland handelte es sich dabei um die Monate von Mai bis Juli. In zwei Bundesländern wurde die behördliche Genehmigung so spät erteilt, dass die Datenerhebung erst zu Beginn des darauffolgenden Schuljahres – ca. vier Wochen nach Schuljahresbeginn – stattfinden konnte. Es wurde daher in diesen Fällen jeweils in der nächsthöheren Klassenstufe erhoben. Zusätzlich wurden die Daten der Erhebung für die Validierung der Tests, die zeitgleich mit den letzten Erhebungen der Normierung stattfand, in die Normstichprobe mit aufgenommen.

Über die Durchführung der Endversionen von LESEN 6-7 und LESEN 8-9 hinaus wurden im Rahmen der Normierung folgende weitere Informationen von den Schülern erfragt: Geschlecht, Geburtsmonat und -jahr, Klassenstufe, Schulart und Muttersprache sowie die Deutschnote im letzten Schulzeugnis und das Vorliegen einer LRS-Diagnose. Durchgeführt wurden alle Testungen von wissenschaftlichen Mitarbeitern von Universitäten oder deren Hilfskräften. Alle Testleiter hatten sich zuvor mit den Tests und den Instruktionen vertraut gemacht.

14.2 Stichprobenanalyse

Im Folgenden ist die vor der Datenauswertung vorgenommene Datenkontrolle dargestellt, die zum Ausschluss einzelner Schüler führte. Anschließend werden die verbleibenden Stichproben beschrieben und die jeweilige Größe der Gesamtstichproben und jeweils verschiedener Substichproben sowie deren Repräsentativität kritisch betrachtet.

Datenkontrolle und Fallausschlüsse. Schüler, bei denen der Testleiter auf dem Testheft vermerkt hatte, dass sie die Instruktion nicht richtig verstanden hatten oder vorzeitig die Testbearbeitung abgebrochen haben, wurden nicht in die Datensätze aufgenommen. Darüber hinaus wurden weit außerhalb der Verteilung liegende Testergebnisse ausgeschlossen, da sie einen hohen Messfehleranteil aufweisen und die Streuung der Messwerte verzerren würden.

Die Prüfung auf „Ausreißerwerte“ geschah pro Klassenstufe und Schulart. Da Leistungstests nicht nach oben verfälscht werden können, wurde lediglich nach Ausreißern in Form extrem schlechter Subtestergebnisse gesucht. Dabei galt ein Subtestergebnis dann als Ausreißer, wenn es mindestens drei Standardabweichungen unter dem Mittelwert der Verteilung lag (vgl. Stevens, 2002, S. 16f.). Für den Subtest TV galt ein Wert zudem erst als Ausreißer, wenn er zusätzlich unter dem Wert lag, den ein Schüler bei zufälligem Ankreuzen, also durch völlig vorwissensfreies Raten erreichen würde. Die Wahrscheinlichkeit, per Zufall die richtige Lösung anzukreuzen, liegt mit jeweils fünf Antwortalternativen pro Item für jedes Item bei 20 %. Würde ein Schüler bei allen Items raten, würde er somit bei LESEN 6-7 im Subtest TV (34 Items) 6,8 Punkte erzielen und bei LESEN 8-9 (38 Items) 7,6 Punkte. Ein genereller Ausschluss von

Schülern mit nicht deutscher Muttersprache oder LRS erfolgte nicht, da die Normen ein möglichst realistisches Abbild der tatsächlichen Verhältnisse darstellen sollten.

Von jenen Schülern, die den Subtest BLK komplett ausfüllten und bei denen das Testheft nicht mit einer Anmerkung des Testleiters versehen war, wurde lediglich ein Schüler der sechsten Gymnasialklasse ausgeschlossen, da sein Wert mehr als drei Standardabweichungen ($SD = 13.87$) unter dem Mittelwert von 61 Punkten aller Schüler der sechsten Gymnasialklassen lag. Mit nur 13 Punkten lag er sogar unterhalb von drei Standardabweichungen vom Mittelwert der sechsten Hauptschulklassen und das, obwohl er angab, Deutsch als Muttersprache und keine LRS-Diagnose zu haben. Bei LESEN 8-9 musste kein Schüler aufgrund eines extrem schlechten Ergebnisses im Subtest BLK ausgeschlossen werden.

Aufgrund ihrer schlechten Ergebnisse im Subtest TV wurden bei LESEN 6-7 zwei Schüler der siebten Gymnasialklasse ausgeschlossen und bei LESEN 8-9 ein Schüler der achten Gymnasialklasse. Diese drei Schüler lagen sowohl mehr als drei Standardabweichungen unter dem Mittelwert als auch unter dem Wert, den sie bei zufälligem Ankreuzen erzielt hätten. Alle drei Schüler gaben an, keine LRS-Diagnose und Deutsch als Muttersprache zu haben. In der siebten Gymnasialklasse erzielten die beiden auffälligen Schüler 4 bzw. 5 von 34 möglichen Punkten, der Schüler in der achten Gymnasialklasse kam auf 7 von 38 möglichen Punkten. Aufgrund ihrer schlechten Ergebnisse im Subtest TV wurden bei LESEN 6-7 zwei Schüler der siebten Gymnasialklasse ausgeschlossen und bei LESEN 8-9 ein Schüler der achten Gymnasialklasse. Diese drei Schüler lagen sowohl mehr als drei Standardabweichungen unter dem Mittelwert als auch unter dem Wert, den sie bei zufälligem Ankreuzen erzielt hätten. Alle drei Schüler gaben an, keine LRS-Diagnose und Deutsch als Muttersprache zu haben. In der siebten Gymnasialklasse erzielten die beiden auffälligen Schüler 4 bzw. 5 von 34 möglichen Punkten, der Schüler in der achten Gymnasialklasse kam auf 7 von 38 möglichen Punkten.

Stichprobenbeschreibung. Nach den eben beschriebenen Fallausschlüssen liegen für LESEN 6-7 Daten von 1 644 Schülern vor, für LESEN 8-9 von 945 Schülern. Beide Tests wurden jeweils in sieben deutschen Bundesländern normiert. Insgesamt nahmen Schulen aus den acht Bundesländern Bayern, Baden-Württemberg, Berlin, Brandenburg, Niedersachsen, Nordrhein-Westfalen, Rheinland-Pfalz und Schleswig-Holstein an der Normierung teil. Die Verteilung der Schüler über die Bundesländer sowie die Klassenstufen und Schularten kann Abbildung 16 entnommen werden. Tabelle 13 enthält für alle Klassenstufen und Schularten die Schülerzahlen sowie Informationen zur Geschlechterverteilung, zur Muttersprache und zum Vorliegen einer LRS-Diagnose. Tabelle 17 im Ergebnisteil zeigt die genauen Stichprobengrößen für die einzelnen Subtests pro Klassenstufe und Schulart.

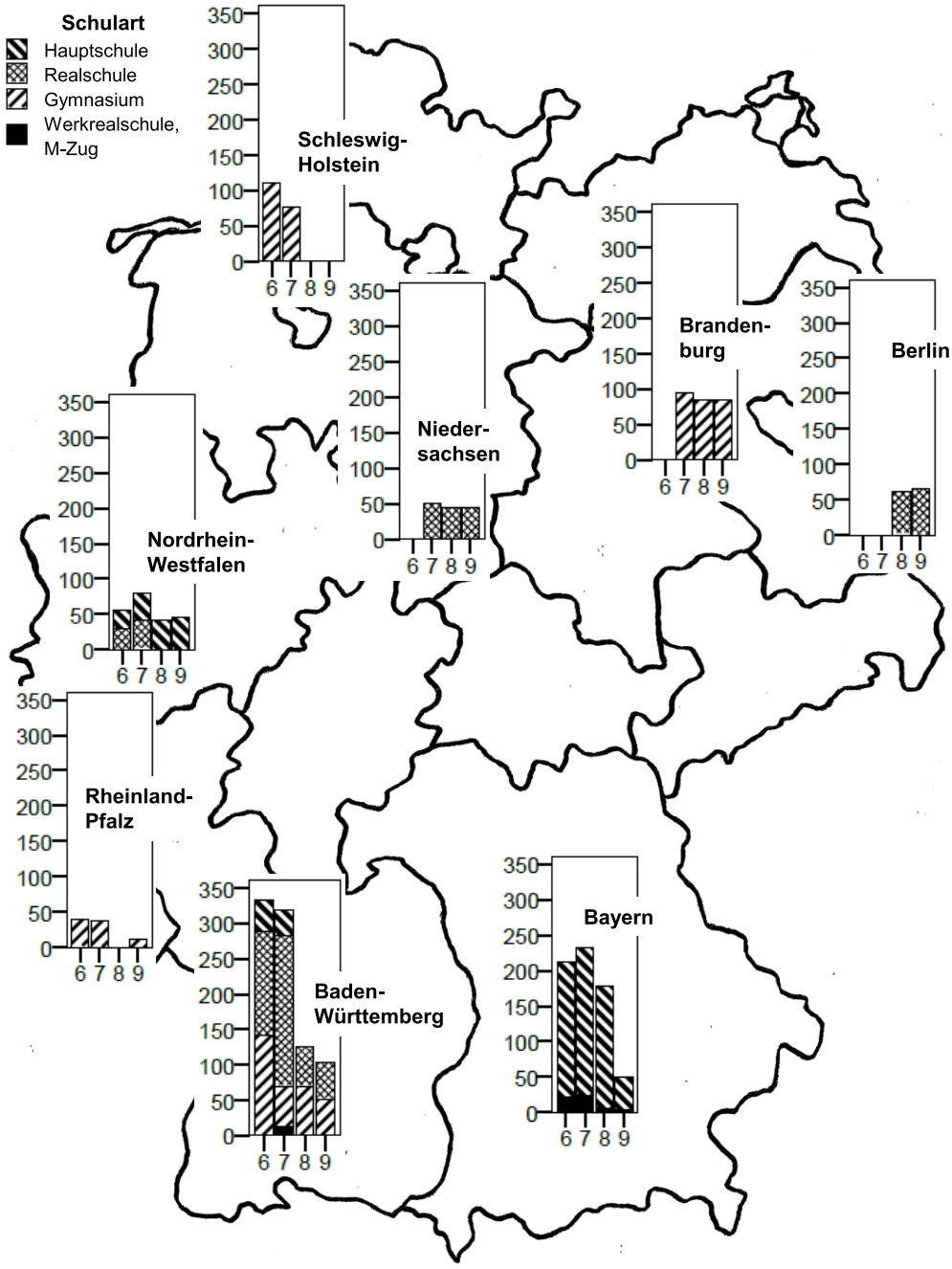


Abbildung 16. Verteilung der Normstichprobe über die Klassenstufen, Schularten und Bundesländer.

Tabelle 13. Schülerzahlen und Informationen zur Zusammensetzung der Normstichprobe.

Klassen-	Schul-		Geschlecht in %		Muttersprache in %			LRS in %	
stufe	art	*N*	m	w	Deutsch	Mehrere	Andere	Nein	Ja
6	HS	263	46	54	61	2	36	81	16
	RS	177	54	46	87	2	10	94	6
	GYM	292	53	47	86	9	2	89	11
	AN	21	52	48	100	0	0	95	5
	Gesamt	753	51	49	78	2	17	87	11
7	HS	283	56	43	65	0	33	84	14
	RS	304	52	47	83	0	16	91	7
	GYM	267	50	50	94	1	5	93	6
	AN	37	41	57	76	0	24	92	8
	Gesamt	891	52	47	81	0	18	89	9
8	HS	215	54	46	67	1	31	87	11
	RS	162	60	38	88	2	7	90	7
	GYM	156	61	39	95	5	0	93	5
	AN	6	29	71	57	0	43	100	0
	Gesamt	539	57	42	81	2	15	90	8
9	HS	92	44	55	74	0	17	88	11
	RS	163	53	46	88	1	9	92	7
	GYM	148	40	60	98	0	2	97	3
	AN	3	67	33	100	0	0	100	0
	Gesamt	406	46	53	88	1	8	93	6

Anmerkungen:
"Mehrere" = Kombination von Deutsch mit einer oder mehreren weiteren Sprache(n)
Abweichungen von 100 % in der Summe sind auf fehlende Angaben zurückzuführen

Fehlende Werte. Aus Tabelle 17 wird ersichtlich, dass einzelne Subtests fehlende Werte aufweisen. Die große Menge an fehlenden Werten für den Subtest TV bei LESEN 6-7 für die sechste Realschulklasse kommt dadurch zustande, dass bei einer Klasse mit 27 Schülern während der Bearbeitung dieses Subtests unerwartet eine Feueralarm-Übung stattfand, weshalb der Test abgebrochen werden musste. Die übrigen fehlenden Werte kommen dadurch zustande, dass die Schüler die Instruktion nicht richtig verstanden hatten oder kurzzeitig abgelenkt waren, was der Testleiter auf dem Testheft notiert hat und weshalb die Daten nicht verwendet wurden. Es wird davon ausgegangen, dass in diesen Fällen der jeweils andere Subtest nicht beeinträchtigt war.

Beurteilung der Substichprobengrößen. Die Substichproben, die sich aus einer Aufteilung nach Klassenstufen sowie nach Klassenstufen und Schularten ergeben, wurden hinsichtlich ihrer Größe bewertet. Hierfür wurden die groben Richtlinien von Wyschkon (2011, S. 87) zum Mindestumfang einer Normstichprobe für die Gewähr-

leistung einer ausreichenden Differenzierung in den Randbereichen der Verteilung herangezogen. Diese Kriterien gelten unter den Voraussetzungen, dass der verwendete Test perfekt reliabel ist, eine Zufallsstichprobe herangezogen wird und in dieser die Leistungsverteilung exakt normalverteilt ist. Da die genannten Voraussetzungen praktisch nie erfüllt sind, werden Wyschkons Kriterien hier als Mindestvoraussetzung angesehen. Wyschkon bezieht sich ausschließlich auf den unteren Leistungsbereich. Da in der vorliegenden Arbeit jedoch auch im oberen Leistungsbereich eine ausreichende Differenzierung gewünscht wird, wurden die Kriterien so formuliert, dass beide Randbereiche der Verteilung eingeschlossen sind. Da von einer Normalverteilung und somit einer symmetrischen Verteilung ausgegangen wird, sollten sowohl für den oberen als auch für den unteren Randbereich die gleichen Bedingungen gelten.

Entsprechend lauten die Kriterien wie folgt:

- Für eine ausreichend präzise Diagnose leichter Auffälligkeiten ($36 < T < 39$ bzw. $61 < T < 64$) sollte die Normstichprobe ca. 100 bis 150 Personen umfassen.
- Für die zuverlässige Identifikation deutlicher Auffälligkeiten ($30 < T < 35$ bzw. $65 < T < 70$) sollte die Normstichprobe ca. 200 bis 250 Personen umfassen.
- Für eine sehr genaue Identifikation extremer Auffälligkeiten ($T < 30$ bzw. $T > 70$) sollte die Normstichprobe mehr als 500 Personen umfassen, wobei für eine akzeptable Schätzung in diesem Bereich die Untergrenze bei einem Stichprobenumfang von ca. 300 Personen liegt.
- Bei einem Stichprobenumfang von weniger als 150 Personen kann in den Extrembereichen nicht differenziert werden.

Die von Wyschkon geforderten Stichprobengrößen gehen etwa mit der Empfehlung von Bühner (2011) konform, mindestens 300 Personen in eine Normstichprobe aufzunehmen.

Tabelle 14 zeigt die Stichprobengrößen für LESEN 6-7 und LESEN 8-9 pro Klassenstufe und Schulart sowie schulartübergreifend. Die zuvor unter „Andere“ aufgeführten Schüler werden hier nicht als separate Schulart einbezogen, da für diese Substichprobe aufgrund der geringen Fallzahlen und mangelnder Repräsentativität keine eigenen Normtabellen erstellt wurden. Daraus ergibt sich auch die Abweichung der Werte in der Spalte „Gesamt“ von der Summe der drei Werte für die angeführten Schularten. Die grauen Abstufungen und die Schriftarten (kursiv, normal, fett) spiegeln die Bewertung der Substichprobengrößen im Hinblick auf die Minimalanforderungen zur Differenzierung in verschiedenen Auffälligkeitsbereichen wider (vgl. Wyschkon, 2011, S. 87). Die schulartübergreifenden Substichproben auf Klassenstufenebene sowie die Substichprobe der Realschüler der siebten Klassenstufe sind demnach ausreichend groß, um eine Differenzierung im Bereich extremer Auffälligkeiten zu ermöglichen.

Die Substichproben der Hauptschüler in den Klassentufen sechs, sieben und acht sowie die Substichproben der Gymnasiasten in den Klassenstufen sechs und sieben sind immerhin noch jeweils ausreichend groß, um eine Differenzierung im Bereich deutlicher Auffälligkeiten zu erlauben. Die Größe der Substichproben der sechsten, achten und neunten Realschulklasse sowie der achten und neunten Gymnasialklasse reichen lediglich für eine Differenzierung im Bereich leichter Auffälligkeiten aus. Die Substichprobengröße der neunten Hauptschulklasse erlaubt keine Differenzierung in den Randbereichen.

Dabei ist zu beachten, dass das Erreichen der Mindeststichprobengröße allein nicht garantiert, dass die Differenzierungen tatsächlich möglich sind. Hierfür muss z. B. auch die Qualität der Items und die Repräsentativität der Stichprobe gewährleistet sein. Die Mindeststichprobengröße sagt lediglich aus, dass die Differenzierung in den entsprechenden Bereichen bei ansonsten optimalen Bedingungen möglich ist. Ob die Voraussetzungen der Repräsentativität und der Normalverteilung für die vorliegende Normstichprobe gegeben sind, wird in den nächsten Abschnitten geprüft.

Tabelle 14. Substichprobengrößen und Beurteilung bezüglich der Differenzierungsfähigkeit für die einzelnen Klassenstufen nach Schulart getrennt sowie schulartübergreifend.

	Hauptschule	Realschule	Gymnasium	Gesamt
6. Klasse	263	*177*	292	**753**
7. Klasse	283	**304**	267	**891**
8. Klasse	215	*162*	*156*	**540**
9. Klasse	92	*163*	*148*	**406**

Anmerkungen:

· weiß/normal: keine Differenzierung in den Randbereichen möglich ($N < 100$)

· hellgrau/kursiv: leichte Auffälligkeiten differenzierbar ($N > 100$)

· mittelgrau/normal: deutliche Auffälligkeiten differenzierbar ($N > 200$)

· dunkelgrau/fett: extreme Auffälligkeiten differenzierbar ($N > 300$)

Repräsentativität der Normstichprobe. Im Gegensatz zu großen, von staatlicher Seite veranlassten Studien wie z. B. PISA, die eine annähernd repräsentative Erhebung durchführen können, ist es in kleineren Studien wie der vorliegenden, die nicht mit den entsprechenden Ressourcen ausgestattet und auf die freiwillige Teilnahme von Schulen, Lehrkräften und Schülern angewiesen sind, praktisch unmöglich, eine repräsentative Stichprobe zu untersuchen. Bei Minderjährigen muss außerdem eine Einverständniserklärung der Eltern vorliegen. Durch die daher nicht zu vermeidende Selbstselektion kann eine deutliche Verzerrung der Stichprobe entstehen (vgl. Wyschkon, 2011, S. 299ff.).

Um die Repräsentativität der Normstichproben für LESEN 6-7 und LESEN 8-9 einschätzen zu können, wurde zunächst überprüft, inwieweit die Normstichprobe in jeder Klassenstufe in wesentlichen Merkmalen von der entsprechenden Zielpopulation

abweicht. In Testmanualen von Leistungstests für Sekundarschüler wird bei der Beurteilung der Repräsentativität der Normstichprobe häufig berichtet, inwieweit die Verteilung folgender Merkmale in der Normstichprobe jener in der Zielpopulation entspricht (Wyschkon, 2011, S. 65): Geschlecht, Bildungsabschlüsse/Berufe der Eltern, Wohnregion (städtisch vs. ländlich) und Schulart. In der vorliegenden Studie wurden bei der Normierung die Art der Wohnregion sowie die Bildungsabschlüsse bzw. Berufe der Eltern nicht erfasst. Williams und Bowman (2002) schlussfolgern in ihrem Literaturüberblick, dass sich bedeutsame Leistungsdiskrepanzen zwischen Kindern aus städtischen und ländlichen Regionen vorwiegend in älteren Studien fanden. In neueren Studien zeigten sich höchstens sehr kleine Unterschiede und diese vor allem bei jungen Kindern, da – so vermuten Williams und Bowman – die formale Schulbildung bis zum Alter von etwa elf Jahren bei den Kindern aus ländlichen Regionen zum Aufschließen auf den Leistungsstand der Kinder städtischer Regionen führt. Somit ist bei LESEN 6-7 und LESEN 8-9 kein Leistungsunterschied zwischen Schülern aus städtischen Regionen und Schülern aus ländlichen Regionen mehr zu erwarten. Die elterlichen Schulabschlüsse und Berufe wurden lediglich im Rahmen der Validierung erhoben (vgl. Kap. 17). Dabei führte jedoch die Erfassung dieser Merkmale über Schülerfragebögen zum Teil zu sehr unzuverlässigen Ergebnissen und vielen fehlenden Angaben. Vom zusätzlichen Einsatz eines Elternfragebogens wurde abgesehen, da befürchtet wurde, dadurch eine noch stärkere Selbstselektion auszulösen.

Die Merkmale Schulart, Geschlecht, Muttersprache und das Vorliegen einer LRS-Diagnose konnten jedoch zur Repräsentativitätsprüfung herangezogen werden. Darüber hinaus wurde die Verteilung der Stichprobe über die Bundesländer deskriptiv betrachtet. Zur Überprüfung der Verteilung der Schüler über die Schularten sowie der Geschlechterverteilung wurden die Angaben des statistischen Bundesamtes für die Sekundarstufe I im Schuljahr 2010/2011 herangezogen (Baumann, Schneider, Vollmar & Wolters, 2012). Dabei muss berücksichtigt werden, dass bei den Angaben des statistischen Bundesamtes nur Werte für die Sekundarstufe I insgesamt angegeben sind. Abweichungen davon in den einzelnen Klassenstufen sind möglich. Zur Einschätzung der Repräsentativität des Anteils an Schülern mit einer anderen Muttersprache als Deutsch sowie zum Vorliegen einer LRS-Diagnose wurden Ergebnisse bzw. Schätzungen anderer Studien herangezogen. Dort, wo zuverlässige Zahlen über die tatsächliche Verteilung vorliegen, wurden Abweichungen per χ^2-Test auf Signifikanz geprüft. Die Voraussetzungen für den Signifikanztest (Mindesthöhe der erwarteten Häufigkeiten von $n = 5$ pro Zelle und Unabhängigkeit der Beobachtungen; vgl. Bortz & Schuster, 2010, S. 141) konnten als gegeben gelten. In Bezug auf die übrigen Merkmale wurde eine rein deskriptive Beurteilung vorgenommen.

Tabelle 15 zeigt eine Gegenüberstellung der Angaben des statistischen Bundesamtes und der Normstichprobe im Hinblick auf die Verteilung der Schüler über die *Schularten*. Da in der Normstichprobe unter „Andere“ nur Schulen mit mehreren Bildungsgängen gefasst wurden, wurden für „Andere“ die Angaben des statistischen Bun-

desamtes zu dieser Schulart herangezogen. Es fällt auf, dass vor allem dieser Bereich stark unterrepräsentiert ist. Hauptschulklassen sind durchweg – in der neunten Klasse leicht und ansonsten zum Teil deutlich – überrepräsentiert. Die Realschüler sind in der sechsten Klasse unter- und ansonsten überrepräsentiert. Gymnasiasten sind durchweg mehr oder weniger stark unterrepräsentiert. χ^2-Tests zeigen, dass die Abweichungen von der erwarteten Verteilung für alle Klassenstufen signifikant sind (Klassenstufen 6-8: $p < .01$; Klassenstufe 9: $p = .01$). Die Überrepräsentation der Hauptschüler und Unterrepräsentation der Gymnasiasten könnte zu der Annahme führen, dass die schulartübergreifenden Normen der Tests insgesamt die Leistung von Schülern eher überschätzen. Die Bedeutsamkeit der Nichtrepräsentativität bezüglich der Verteilung der Schüler über die verschiedenen Schularten hängt dabei auch vom Ausmaß der Leistungsunterschiede zwischen den Schularten ab. Bisherige Studien fanden deutliche Unterschiede zwischen den Schularten (s. Kap. 6).

Darüber hinaus ist bei der Bewertung der Ergebnisse zu berücksichtigen, dass die Freiwilligkeit der Teilnahme dazu führte, dass in einigen Klassen zahlreiche Schüler die Teilnahme verweigerten bzw. die Eltern der Teilnahme ihres Kindes nicht zustimmten. Vorherige Untersuchungen zeigten, dass gerade Schüler, von denen erwartet würde, dass sie schwache Leistungen zeigen, nicht an Normierungsstudien teilnehmen (vgl. z. B. Wyschkon, 2011, S. 15). Auch dies lässt vermuten, dass LESEN 6-7 und LESEN 8-9 die Leistungen der Schüler eher überschätzen. Weiter ist jedoch auch die Verteilung der Schüler der Normstichproben über die Bundesländer zu berücksichtigen, die später geschildert wird.

Tabelle 15. Gegenüberstellung der Angaben (in %) des Statistischen Bundesamtes zur Schülerverteilung über die Schularten in der Sekundarstufe I und den Normstichproben von LESEN 6-7 und LESEN 8-9.

	HS	RS	GYM	AN
Statistisches Bundesamt, Sek. I	19	31	40	10
LESEN 6-7, 6. Klasse	35	23	39	3
LESEN 6-7, 7. Klasse	32	34	30	1
LESEN 8-9, 8. Klasse	40	30	29	4
LESEN 8-9, 9. Klasse	23	40	36	1

Als nächstes wird die Repräsentativität des *Geschlechterverhältnisses* betrachtet. Tabelle 16 zeigt eine diesbezügliche Gegenüberstellung der Angaben des Statistischen Bundesamtes für die Sekundarstufe I mit der Normstichprobe von LESEN 6-7 und LESEN 8-9. In den Klassenstufen sechs und sieben stimmt das Geschlechterverhältnis gut mit den Angaben des statistischen Bundesamtes (2012) überein. In der achten Klasse sind die Jungen bei LESEN 8-9 mit einem Anteil von 58 % deutlich überrepräsentiert und in der neunten Klasse mit einem Anteil von 47 % unterrepräsentiert. χ^2-Tests zeigten, dass die Abweichungen von der erwarteten Geschlechterverteilung nur

für die Klassenstufen acht und neun signifikant ausfallen (6. Klasse: $p = .79$, 7. Klasse $p = .81$, 8. Klasse: $p < .01$, 9. Klasse: $p = .04$). Ob diese Abweichungen die Güte der Normstichprobe deutlich beeinträchtigen, hängt auch hier von Leistungsunterschieden zwischen den Geschlechtern ab. Bisherige Studien fanden zwar meist signifikante Geschlechterunterschiede in der Leseleistung, die diesbezüglichen Effektstärken waren jedoch meist gering und praktisch nicht bedeutsam (s. Kap. 6).

Tabelle 16. Gegenüberstellung der Angaben (in %) des statistischen Bundesamtes zur Geschlechterverteilung in der Sekundarstufe I und der Normstichprobe von LESEN 6-7 und LESEN 8-9.

	m	w
Statistisches Bundesamt, Sek I	52	48
LESEN 6-7, 6. Klasse	51	49
LESEN 6-7, 7. Klasse	53	47
LESEN 8-9, 8. Klasse	58	42
LESEN 8-9, 9. Klasse	47	53

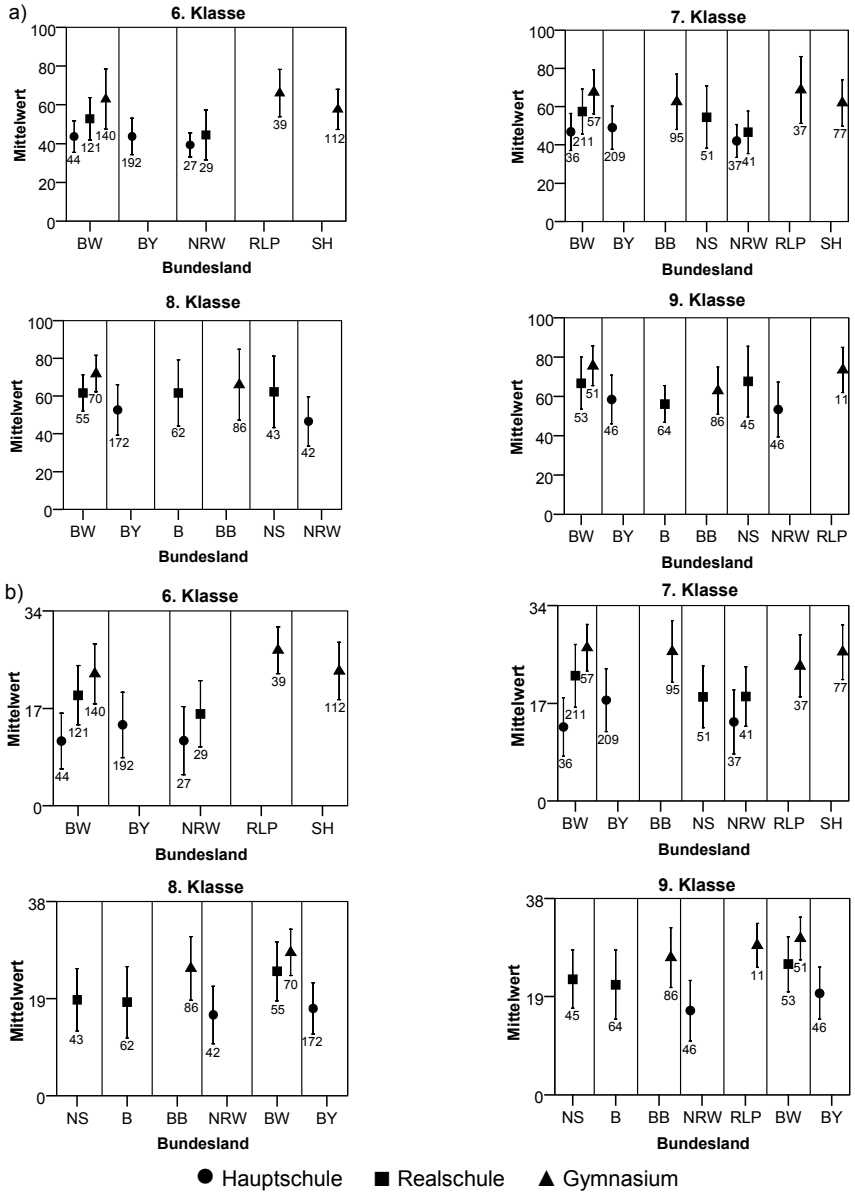
Im Hinblick auf *Muttersprache* ist es schwierig, zuverlässige Angaben über die aktuelle Lage an deutschen Regelschulen zu bekommen. Sowohl PISA als auch das Statistische Bundesamt differenzieren nicht nach der Muttersprache, sondern nach dem Migrationshintergrund, wobei der Migrationsstatus über das eigene Geburtsland des Schülers und das Geburtsland der Eltern bestimmt wird. DESI hingegen erhob die Erstsprache der Neuntklässler, wobei sich ein enger Zusammenhang zwischen Migrationshintergrund und Erstsprache zeigte (DESI-Konsortium, 2006, S. 22). Bei DESI gaben 13 % der Schüler an, ausschließlich eine andere Sprache als Deutsch in ihrer Familie gelernt zu haben, und 6 % gaben an, Deutsch und eine weitere Sprache in ihrer Familie gelernt zu haben. In der vorliegenden Normstichprobe gaben bei LESEN 6-7 knapp 18 % der Schüler an, eine andere Muttersprache als Deutsch zu haben und etwas mehr als 1 % der Schüler gab an, Deutsch und eine weitere Sprache als Muttersprachen zu haben. Bei LESEN 8-9 gaben ca. 12 % der Schüler an, eine andere Sprache als Deutsch als Muttersprache zu haben, während ca. 2 % mehrere Sprachen angaben. Somit stimmen die Angaben für die vorliegenden Normstichproben zumindest mit den Angaben relativ gut überein, die bei DESI für Neuntklässler in Bezug auf den Anteil an Schülern mit einer anderen Muttersprache als Deutsch gemacht werden. Der Anteil mehrsprachiger Schüler ist in den vorliegenden Normstichproben geringer.

In Bezug auf die Diagnose von *LRS* gaben bei LESEN 6-7 insgesamt 10 % der Schüler an, dass bei ihnen LRS diagnostiziert wurde, bei LESEN 8-9 lag der Anteil bei 7 %. Inwiefern diese Zahlen repräsentativ sind, ist schwer einzuschätzen, da die Prävalenzzahlen abhängig von den Definitions- und Diagnosekriterien stark schwanken (Berger, 2010, S. 19). Unter Bezug auf die Diagnosekriterien des ICD-10 und DSM-IV-TR kamen verschiedene Studien zu dem Ergebnis, dass 3 % bis 9 % der Kinder und Jugend-

lichen in Deutschland Schwierigkeiten mit der Schriftsprache haben (Schulte-Körne, 2004; Hasselhorn & Schuchardt, 2006; H. Marx & Reinhold, 2010; Strehlow & Haffner, 2002). Insgesamt scheinen die Werte von 6 % bis 11 % in den einzelnen Klassenstufen der Normstichprobe von LESEN 6-7 und LESEN 8-9 (s. Tab. 13) somit realistisch. Natürlich ist zu berücksichtigen, dass es sich dabei ausschließlich um Selbstauskünfte der Schüler handelt, und dass es Hinweise darauf gibt, dass besonders schwache Schüler nicht an Untersuchungen teilnehmen (vgl. Wyschkon, 2011, S. 15).

Bezüglich der Verteilung der Schüler über die *Bundesländer* fällt auf, dass in nördlichen Bundesländern keine Hauptschuldaten erhoben wurden (s. Abb. 16). Auch die übrigen Schularten sind nicht gleichmäßig über Deutschland verteilt. Um herauszufinden, wie sehr dies die Güte der Normstichproben beeinträchtigt, ist es nötig, zu analysieren, ob zwischen den Bundesländern bedeutsame Unterschiede bestehen. In der Bundesländervergleichsstudie des IQB aus dem Jahr 2009 wurden signifikante Unterschiede zwischen den Bundesländern gefunden (Schipolowski & Böhme, 2010, s. auch Kap. 6). In der Normstichprobe von LESEN 6-7 und LESEN 8-9 sind die Bundesländer Bayern und Baden-Württemberg vertreten, die in der genannten Studie signifikant überdurchschnittliche Ergebnisse erzielten. Ebenfalls vertreten sind die Bundesländer Berlin und Brandenburg, die in derselben Studie beim Lesen unterdurchschnittliche Ergebnisse erzielten. Daher ist davon auszugehen, dass auch bei der Normierung von LESEN 6-7 und LESEN 8-9 entsprechende Leistungsunterschiede zwischen den Bundesländern auftreten. Da jedoch nur ein Teil der Bundesländer in die Normierung einbezogen war und dies in sehr ungleichen Verhältnissen (z. T. nur eine Schulart und/oder eine Klasse pro Klassenstufe), tritt einerseits eine starke Konfundierung von Klassen-/Schulzugehörigkeit und Bundeslandzugehörigkeit auf, was bei der Ergebnisinterpretation zu berücksichtigen ist. Andererseits wären bei einer inferenzstatistischen Prüfung die Zellen sehr ungleich besetzt und zum Teil unbesetzt, sodass auf eine inferenzstatistische Prüfung verzichtet und lediglich deskriptiv analysiert wurde: Abbildung 17 zeigt, dass zum einen innerhalb einer Schulart große Unterschiede zwischen den Bundesländern bestehen und zum anderen die Unterschiede zwischen Schularten unterschiedlicher Bundesländer gering sind. So liegt z. B. beim Subtest TV in der siebten Klassenstufe der Mittelwert der bayerischen Hauptschulen nahe am Mittelwert der nordrhein-westfälischen und niedersächsischen Realschulen, und der Mittelwert der baden-württembergischen Realschule liegt nahe am rheinland-pfälzischen Gymnasium. Es zeigen sich also in der Normierung von LESEN 6-7 und LESEN 8-9 tendenzielle Unterschiede zwischen den Bundesländern, große Unterschiede sind jedoch nur vereinzelt zu erkennen (nach Bundesländern aufgeteilte deskriptive Statistiken s. Anhang B, Tab. 41).

Aufgrund der eben beschriebenen Einschränkungen können die Normstichproben nicht in Bezug auf alle relevanten Aspekte als repräsentativ bezeichnet werden. Insbesondere was die Verteilung über die Bundesländer und über die Schularten angeht, ist



BW = Baden-Württemberg, BY = Bayern, NRW = Nordrhein-Westfalen, RLP = Rheinland-Pfalz, SH = Schleswig-Holstein, BB = Brandenburg, B = Berlin, NS = Niedersachsen

Abbildung 17. Bundesländervergleich für alle Schularten nach Klassenstufen getrennt; a: Subtest BLK (jeweils max. 100 Punkte); b: Subtest TV (6. und 7. Klasse max. 34 Punkte, 8. und 9. Klasse max. 38 Punkte). Fehlerbalken: ± 1 Standardabweichung; Zahlen unterhalb der Fehlerbalken: Schülerzahlen.

Repräsentativität nicht gegeben. Realistisch erscheinen hingegen die Zahlen bezüglich des Geschlechterverhältnisses (zumindest für die Klassenstufen sechs und sieben), des Anteils an Schülern mit einer anderen Muttersprache als Deutsch und des Anteils an Schülern mit einer LRS-Diagnose.

14.3 Deskriptive Statistik und Verteilungen

Die deskriptiven Werte jedes Subtests wurden für die Gesamtstichproben und für die Substichproben nach Klassenstufen getrennt sowie innerhalb der Klassenstufen nach Schulart getrennt betrachtet (s. Tab. 17). Zur Beurteilung der Rohwertverteilungen wurde zusätzlich zur Betrachtung der Histogramme mit Normalverteilungskurve (s. Abb. 18) jeweils ein Kolmogorov-Smirnov-Test (KS-Test, Prüfgröße *Z*) auf Normalverteilung berechnet (s. Tab. 17). Die Signifikanzprüfung der KS-Tests erfolgte mit einem α-Niveau von $.20$, da die Wahrscheinlichkeit für eine fälschliche Annahme einer Normalverteilung gering gehalten werden sollte.

Es zeigt sich hypothesenkonform in beiden Subtests ein Anstieg der Rohwerte von Klassenstufe zu Klassenstufe sowie von der Hauptschule über die Realschule zum Gymnasium. Die Leistung der unter „Andere“ zusammengefassten Schüler aus den Werkrealschulen und M-Zug-Klassen liegt erwartungsgemäß größtenteils zwischen den Haupt- und Realschülern, in der achten Klasse liegen sie in ihrer Leistung im Subtest BLK sogar über den Gymnasiasten (allerdings befinden sich dort lediglich sechs Schüler in der Kategorie „Andere“). Der Mittelwert des Gesamtergebnisses liegt in jeder Klassenstufe gerundet bei 100 und die Standardabweichung bei ungefähr 15, da sich das Gesamtergebnis jeweils aus der Summe der T-Werte der Subtests ergibt.

Für die unter „Andere“ zusammengefassten Schularten wurde aufgrund der kleinen Stichproben kein KS-Test berechnet. Für diese Kategorie wurden auch keine eigenen Normtabellen erstellt. Der KS-Test auf Normalverteilung für die Gesamtstichproben beider Tests fällt für beide Subtests signifikant aus (LESEN 6-7: $KS\text{-}Z = 3.10$ für BLK, $KS\text{-}Z = 2.71$ für TV; LESEN 8-9: $KS\text{-}Z = 2.38$ für BLK, $KS\text{-}Z = 2.08$ für TV; alle $p < .01$) und für das Gesamtergebnis jeweils nicht (LESEN 6-7: $KS\text{-}Z = 0.81$, $p = .53$; LESEN 8-9 $KS\text{-}Z = 1.04$, $p = .23$). Ersteres war aufgrund der Stichprobengröße zu erwarten, letzteres aufgrund der Tatsache, dass sich das Gesamtergebnis aus der Summe der genormten Subtestergebnisse ergibt. Auch auf Klassenstufen- und Schulartebene wird die Abweichung von der Normalverteilung für die Subtests zu einem großen Teil signifikant, für das Gesamtergebnis nie.

Tabelle 17. Deskriptive Statistik und Kolmogorov-Smirnov-*Z* (*KS-Z*) für beide Subtests (BLK und TV) auf Basis der Normdaten nach Klassenstufe und Schulart getrennt.

LESEN 6-7									
		BLK				TV			
Klasse	Schulart	*N*	*M*	*SD*	*KS-Z*	*N*	*M*	*SD*	*KS-Z*
6	HS	263	43.24	8.99	1.13*	263	13.36	5.71	1.31*
	RS	177	51.49	11.64	1.02	150	18.67	5.45	0.98
	GYM	291	61.34	13.60	1.49*	292	23.76	5.18	1.57*
	AN	21	49.62	7.39	–	21	18.76	5.31	–
	Σ	752	52.37	13.92	2.22*	726	18.80	7.08	1.91*
7	HS	282	47.82	10.96	1.15*	283	16.46	5.75	1.04*
	RS	303	55.49	13.00	1.25*	304	20.68	5.64	1.42
	GYM	267	64.33	13.87	1.82*	266	25.80	5.02	1.91*
	AN	37	51.32	11.75	–	37	19.05	6.25	–
	Σ	889	55.54	14.20	2.16*	890	20.80	6.64	1.94*
LESEN 8-9									
		BLK				TV			
Klasse	Schulart	*N*	*M*	*SD*	*KS-Z*	*N*	*M*	*SD*	*KS-Z*
8	HS	214	51.51	13.45	1.26*	215	16.84	5.17	1.11*
	RS	160	61.85	15.62	1.71*	160	20.52	6.90	0.73
	GYM	156	68.63	15.55	1.00	156	26.33	5.73	1.71*
	AN	6	71.50	13.05	–	6	22.00	3.23	–
	Σ	536	59.80	16.41	2.09*	537	20.75	7.04	1.56*
9	HS	92	55.97	13.38	1.23*	92	17.97	5.68	0.82
	RS	162	62.85	14.41	1.43*	163	22.90	6.20	1.19*
	GYM	148	68.08	12.82	0.89	148	27.98	5.43	1.57*
	AN	3	58.67	5.51	–	3	25.67	2.31	–
	Σ	405	63.17	14.28	1.50*	406	23.65	6.91	1.61*

Anmerkung: *: $p < .20$

Obwohl die Rohwertverteilungen zum Teil leicht schief ausfallen, scheinen die Tests dennoch im gesamten Leistungsspektrum zu differenzieren. Bei LESEN 6-7 erreicht im Subtest BLK kein Schüler die volle Punktzahl von 100, und im Subtest TV erreicht nur ein Schüler die volle Punktzahl von 34. Bei LESEN 8-9 erreicht ein Schüler im Subtest BLK die volle Punktzahl von 100 und im Subtest TV erreicht kein Schüler die volle Punktzahl von 38. Es liegt also kein Deckeneffekt vor. Die Abbildungen für alle Rohwertverteilungen pro Klassenstufe nach Schularten aufgeteilt befinden sich in Anhang B (Abb. 24 bis 29).

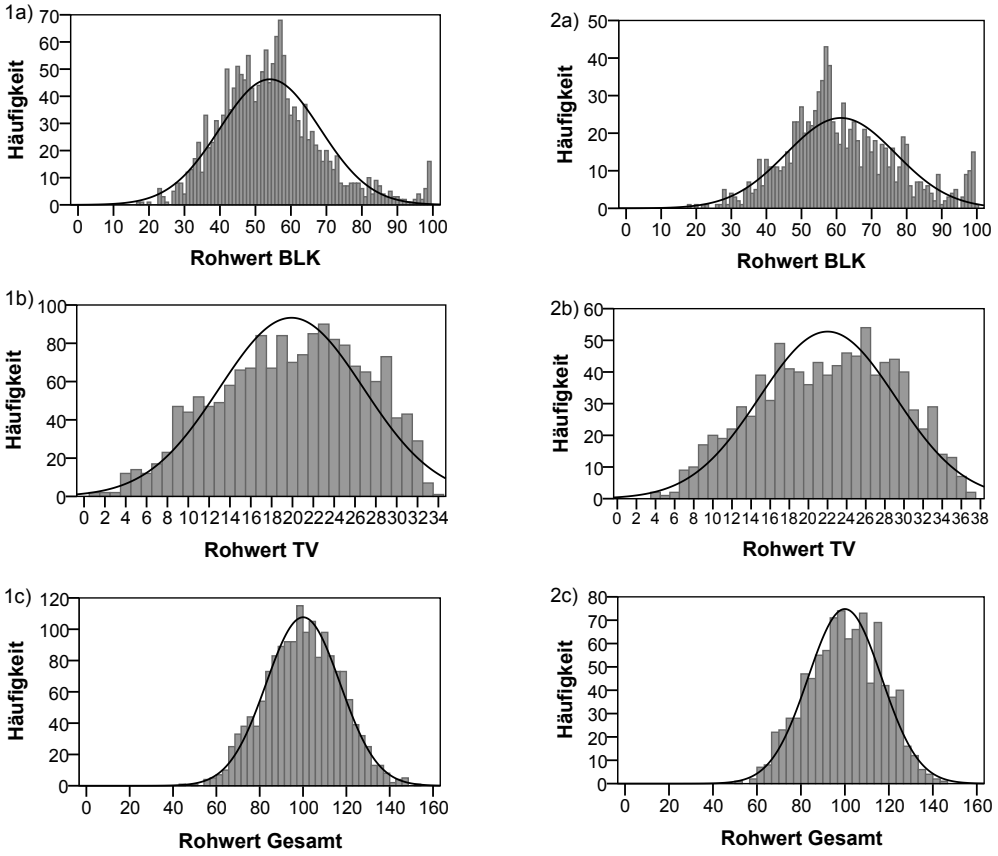


Abbildung 18. Verteilung der Rohwerte für LESEN 6-7 (1) und LESEN 8-9 (2) jeweils für den Subtest BLK (a), den Subtest TV (b) und das Gesamtergebnis (c) auf Basis der Daten der Normstichprobe.

14.4 Erstellung der Normtabellen

Wie bereits erläutert, wurden Normtabellen mit Prozenträngen, T-Werten und T-Wertbändern sowie Personenparametern erstellt. Obwohl sich zum Teil Geschlechterunterschiede und Unterschiede zwischen Schülern mit Deutsch als Muttersprache und Schülern mit einer anderen Muttersprache als Deutsch ergaben (s. Kapitel 17), wurde darauf verzichtet, getrennte Normtabellen für andere Merkmale als Klassenstufe und Schulart zu erstellen, da in den genannten Fällen objektive Leistungsunterschiede vorliegen und sich sowohl die Anforderungen in der Schule als auch im Alltag und im späteren Beruf nicht zwischen diesen Subgruppen unterscheiden. Diejenigen Schüler, die in den Tests niedrige Werte erzielen, sollten unabhängig vom Geschlecht und der

Muttersprache besondere Förderung erfahren. Zudem ist zu berücksichtigen, dass die Unterschiede im Hinblick auf das Geschlecht zwar zum Teil statistisch signifikant ausfallen, die Effektstärken jedoch gering sind.

Prozentränge, T-Werte und T-Wertbänder. Prozentränge werden angegeben, da diese besonders leicht zu interpretieren sind, was vor allem für nicht fundiert psychologisch-diagnostisch ausgebildete Anwender (z. B. Lehrkräfte) die Ergebnisinterpretation erleichtern soll. Zudem sind sie auch für schief verteilte Messwerte verwendbar. Allerdings haben Prozentränge den Nachteil, dass sie nur Ordinalskalenniveau aufweisen und keine Berechnung von Mittelwerten oder Differenzen erlauben.

Aufgrund dieses Nachteils werden zusätzlich T-Werte angegeben. Da sich die Rohwerte nicht bei allen Schularten und Klassenstufen als normalverteilt erwiesen, wurde zur Berechnung der T-Werte zunächst eine Normalrangtransformation durchgeführt. T-Werte besitzen Intervallskalenniveau und ermöglichen die Berechnung von Konfidenzintervallen (vgl. Bühner, 2011, S. 264). Sie wurden anhand der von Amelang und Schmidt-Atzert (2006, S. 51) angegebenen Formel berechnet. In den Testmanualen für LESEN 6-7 und LESEN 8-9 sind T-Wertbänder mit einer Sicherheitswahrscheinlichkeit von 90 % angegeben. Sollte eine andere Sicherheitswahrscheinlichkeit benötigt werden, ist eine Berechnung des entsprechenden Konfidenzintervalls anhand der in den Testmanualen abgedruckten Formel möglich. In den Normtabellen wurde auf die Angabe von Nachkommastellen verzichtet, um nicht eine Messgenauigkeit vorzutäuschen, die die Tests nicht gewährleisten können.

Profildiagramm und Diskrepanzanalyse. Die Leistungen eines Schülers können für einen Überblick als Profildiagramm dargestellt werden, was die Ermittlung von Stärken und Schwächen erleichtert und für die Auswahl individueller Fördermaßnahmen nützlich sein kann. Hierfür steht auf dem Auswertungsbogen eine entsprechende Vorlage zur Verfügung, in die die Werte eines Schülers eingetragen werden können. Abbildung 19 zeigt ein Beispiel eines Profildiagramms für einen Schüler der achten Hauptschulklasse. Der Schüler erzielte im Subtest BLK eine Punktzahl von 40, was einem T-Wert von 41 (T-Wertband 35-51) entspricht. Im Subtest TV erzielte er eine Punktzahl von 22, was einem T-Wert von 59 (T-Wertband 50-63) entspricht. Durch Aufsummieren der T-Werte erhält man das Gesamtergebnis von 100, das wiederum einem T-Wert von 50 (T-Wertband 42-58) entspricht. Der graue Bereich kennzeichnet den Durchschnittsbereich ($M \pm 1\,SD$). Die T-Wertbänder sind in Form von Fehlerbalken eingetragen.

Um festzustellen, ob sich innerhalb des Profils die Leistungen in den Subtests signifikant voneinander unterscheiden, wird vom T-Wert des Subtests BLK der T-Wert des Subtests TV abgezogen. Ob die Differenz außergewöhnlich groß ist und ob es sich um einen signifikanten Unterschied handelt, kann in einer entsprechenden Tabelle im Anhang des Manuals nachgeschlagen werden. Den T-Wert-Differenzen sind dort

die entsprechenden kumulierten Häufigkeiten zugeordnet, wobei jeweils unterschieden werden muss, ob die Differenz negativ (die Teilleistung im Subtest TV ist höher als die Teilleistung im Subtest BLK) oder positiv (die Teilleistung im Subtest BLK ist höher als die Teilleistung im Subtest TV) ausfällt. Die Signifikanzgrenze wurde anhand einer Formel von Amelang und Schmidt-Atzert (2006, S. 54) auf der Basis der Reliabilitätskoeffizienten ermittelt. Im Fall einer Sicherheitswahrscheinlichkeit von 90 % liegt der kritische Wert für den Unterschied der beiden Subtests sowohl bei LESEN 6-7 als auch bei LESEN 8-9 gerundet bei 10 Punkten. D. h. alle Differenzen, die größer als 10 sind, sind als bedeutsam zu bewerten. Nicht signifikante Differenzen sind in der Tabelle im Anhang der Testmanuale entsprechend markiert. Sollte ein anderes Signifikanzniveau gewünscht werden, kann eine im Manual angegebene Formel verwendet werden, um die entsprechende kritische T-Wert-Differenz zu ermitteln.

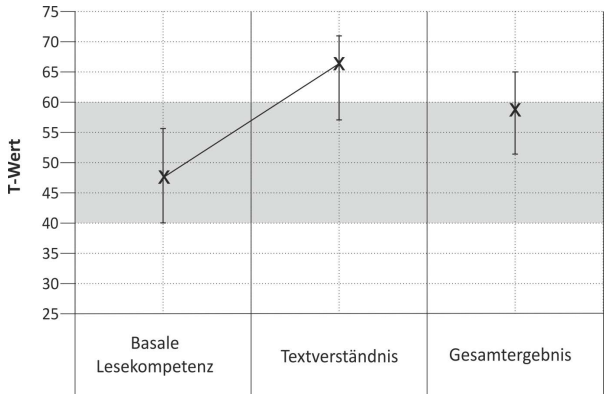


Abbildung 19. Beispiel eines Profildiagramms in einem Auswertungsbogen.

Personenparameter. Für den Subtest TV werden für fortgeschrittene Anwender, die die erweiterten Möglichkeiten der IRT nutzen möchten, zusätzlich Personenparameter mit den entsprechenden Konfidenzintervallen angegeben (Interpretation von Personenparametern s. 5.3.2.2). Jeder im Subtest TV erreichbaren Punktzahl lässt sich ein Personenparameter zuordnen. Die im Manual angegebenen Konfidenzintervalle beziehen sich auf eine Sicherheitswahrscheinlichkeit von 90 % und wurden anhand einer bei Bühner (2011, S. 257) angegebenen Formel auf Basis der Standardfehler berechnet. Auch hier ist die Formel in den Manualen von LESEN 6-7 und LESEN 8-9 abgedruckt, sodass bei Bedarf auch für andere Sicherheitswahrscheinlichkeiten Konfidenzintervalle berechnet werden können.

14.5 Diskussion

Im Folgenden werden das methodische Vorgehen und die Ergebnisse der Normierung diskutiert. Dabei wird zunächst auf die Stichprobengröße und ihre Repräsentativität eingegangen. Anschließend werden die Ergebnisse der deskriptiven Datenanalyse und zuletzt die Erstellung der Normtabellen kritisch betrachtet.

Stichprobengröße. Mit 1 644 Schülern für LESEN 6-7 und 945 Schülern für LESEN 8-9 konnten zufriedenstellende Gesamtstichprobengrößen erreicht werden. Auch bei einer Aufsplittung nach Klassenstufen können die Substichproben noch als ausreichend groß gelten. In allen Klassenstufen wird z. B. die von Bühner (2011) genannte Minimalgrenze von 300 Personen für Normstichproben weit überschritten. Zusätzlich aufgeteilt nach Schulart werden die Stichproben jedoch zum Teil recht klein. Vor allem in der neunten Klassenstufe der Hauptschule werden nicht einmal 100 Schüler erreicht. Dies ist darauf zurückzuführen, dass sich die Schüler der neunten Hauptschulklasse in ihrem Abschlussjahr befanden, weshalb es in diesem Fall besonders schwierig war, eine große Schülerzahl für die Teilnahme zu gewinnen. Die jeweiligen Stichprobengrößen sind in den Testmanualen angegeben und sollten bei der Interpretation der Normwerte berücksichtigt werden.

Repräsentativität. Im Hinblick auf die Repräsentativität der vorliegenden Normstichproben ist zunächst positiv zu bewerten, dass die Zahlen bezüglich des Anteils an Schülern mit einer anderen Muttersprache als Deutsch sowie des Anteils an Schülern mit Vorliegen einer LRS-Diagnose relativ gut mit den Angaben aus anderen Studien übereinstimmen. Weiter übertrifft die Anzahl von jeweils sieben (total acht) in der Normstichprobe vertretenen Bundesländern den Umfang der Normstichproben vieler anderer Schultests (vgl. Kap. 5.3.4). Negativ zu bewerten ist hingegen, dass die ursprünglich geplante gleichmäßige Verteilung der Klassenstufen und Schularten über die Fläche Deutschlands nicht eingehalten werden konnte, da in einigen Bundesländern aufgrund von Umstrukturierungen im Schulsystem nur eine oder zwei bestimmte Schularten und/oder Klassenstufen für die Datenerhebung zur Verfügung standen. Zudem ist zu bemängeln, dass nichts über die Anzahl und Beweggründe der Nichtteilnehmer bekannt ist. Aus praktischen Gründen wurde darauf verzichtet, genau zu erfassen, wie viele der angefragten Schulen, Lehrkräfte und Schüler eine Teilnahme ablehnten.

Die nicht repräsentative Verteilung der Schüler der Normstichprobe über die Bundesländer sowie auch über die Schularten ist durchaus kritisch zu sehen, da aus vorherigen Studien bekannt ist, dass zwischen diesen Subgruppen bedeutsame Unterschiede im Hinblick auf die Leseleistung bestehen. Hinweise darauf, wie stark die Abweichungen der Normstichproben von den Zielpopulationen im Hinblick auf die Schularten die Qualität der Normen beeinträchtigen, werden sich mithilfe inferenzstatistischer Prüfungen auf Unterschiede zwischen den entsprechenden Gruppen in

Kapitel 17 ergeben. Die Unterschiede zwischen den Bundesländern wurden im vorliegenden Kapitel deskriptiv analysiert. Dabei ergab sich eine gewisse Heterogenität zwischen den Bundesländern. Aus diesem Grund müsste, um eine repräsentative Stichprobe zu erreichen, tatsächlich aus allen Bundesländern eine ausreichend große und über alle jeweils existierenden Schularten repräsentativ verteilte Stichprobe gezogen werden. Dabei ist zu berücksichtigen, dass sich die Verteilung der Schülerschaft über die verschiedenen Schularten zwischen den Bundesländern erheblich unterscheidet. Während beispielsweise im Schuljahr 2010/2011 in Hamburg mehr als 40 % der Schüler ein Gymnasium besuchten, traf dies in Brandenburg nur auf etwa 25 % zu. Zugleich müsste die Größe der jeweiligen Stichprobe aus einem Bundesland dem Anteil des Bundeslandes an der gesamtdeutschen Bevölkerung der entsprechenden Altersgruppe entsprechen, um Repräsentativität zu gewährleisten. Weiterhin wären bei einer nicht vollständig zufälligen Ziehung zur Erreichung von Repräsentativität weitere mit dem Leseverständnis in Zusammenhang stehende Aspekte, wie z. B. der Anteil an Schülern mit LRS und Migrationshintergrund zu kontrollieren. All diese Aspekte verdeutlichen, dass es äußerst schwierig bis sogar unmöglich sein dürfte, eine tatsächlich repräsentative Stichprobe zu rekrutieren.

Erschwerend kam bei der Datenerhebung für die Normierung von LESEN 6-7 und LESEN 8-9 hinzu, dass sich die Schulsysteme in einigen Bundesländern gerade in einem grundlegenden Umbau befanden. Selbst bei PISA erwies sich bei der Erhebung im Jahr 2009 eine Unterteilung nach Schularten über die Bundesländergrenzen hinweg als schwierig, da schulstrukturelle Änderungen der letzten Jahre in den einzelnen deutschen Bundesländern zum Teil zu neuen Differenzierungen in den Schularten führten (Naumann et al., 2010).

Es gibt verschiedene Lösungsansätze, wie bei einer nicht repräsentativen Datenerhebung vorgegangen werden kann, um nachträglich die Repräsentativität der Normdaten zu erhöhen (vgl. z. B. Wyschkon, 2011, S. 59ff.). Dazu gehören etwa eine Gewichtung entsprechend eines bedeutenden und nicht repräsentativ erfassten Merkmals oder etwa das zufällige Ziehen von Fällen aus der Gesamtstichprobe, sodass die Substichproben das gleiche Verhältnis aufweisen wie in der Gesamtpopulation. Bei einer Gewichtung oder einer Zufallsziehung können jedoch immer nur einige wenige Merkmale Berücksichtigung finden, wodurch neue Verzerrungen hinsichtlich anderer Merkmale entstehen können. Eine Zufallsziehung aus der vorhandenen Stichprobe führt zudem möglicherweise zu einem großen Verlust an Stichprobengröße.

In Kapitel 7 der vorliegenden Arbeit wurde bereits deutlich, dass Leseverständnis durch zahlreiche Faktoren beeinflusst wird, die nicht alle erhoben wurden und auch unmöglich alle für eine nachträgliche Auswahl oder Gewichtung herangezogen werden können. Laut Alt und Bien (1994) sind die Folgen von Gewichtungen selten vorhersagbar, und eine Gewichtung erscheint nur dann mit ausreichend hoher Wahrscheinlichkeit sinnvoll, wenn die Korrelation des Gewichtungsmerkmals mit dem Zielkriterium über $r = .70$ liegt. Meist führt eine Gewichtung jedoch zu einer Verschlech-

terung der Schätzung. Im vorliegenden Fall liegt in keiner Klassenstufe die Korrelation zwischen Schulart und einem Subtestergebnis oder dem Gesamttestergebnis über $r = .70$. Eine Gewichtung entsprechend der Schulartenanteile in der Gesamtpopulation ist demnach wenig sinnvoll. Weiter ist davon auszugehen, dass eine Gewichtung der Ergebnisse der teilnehmenden Schüler aus einigen Bundesländern nicht die fehlenden Teilnehmer aus anderen Bundesländern ersetzen könnte, da diese sich zum einen in der Lesekompetenz und zum anderen in vielen anderen Merkmalen von den teilnehmenden Schülern unterscheiden dürften.

Aus den genannten Gründen und weil die Vorgehensweisen zu nachträglicher Erhöhung der Repräsentativität die Probleme nicht ausreichender Vergleichbarkeit der Leistungen von Schülern innerhalb einer Schulart, die Unterschiede zwischen den Bundesländern sowie die aktuellen Veränderungen der gesamten Schulsysteme der Bundesländer nicht ausgeglichen hätte, wurde auf einen Versuch einer nachträglichen Erhöhung der Repräsentativität verzichtet. Für die Erstellung der schulartübergreifenden Normen pro Klassenstufe wurden alle jeweils verfügbaren Daten zusammengefasst. Dieses Vorgehen führt zumindest nicht zu einem Verlust an Stichprobengröße.

Die Repräsentativitätsproblematik wird allgemein kontrovers diskutiert. Dabei wird von vielen Forschern grundsätzlich bestritten, dass es überhaupt möglich ist, eine repräsentative Stichprobe zu ziehen. Darüber hinaus herrscht nur schon bei der Definition des Begriffs Repräsentativität Unklarheit; die Definition ist nicht einheitlich, was Wyschkon (2011, S. 36ff.) ausführlich darstellt. Weiter werden in der Literatur keine genauen Angaben dazu gemacht, wie die Repräsentativität am besten überprüft werden kann und inwieweit Abweichungen akzeptabel sind. Es wird lediglich angenommen, dass eine hohe Übereinstimmung in den Randsummen und den verbundenen Beziehungen zwischen Stichprobe und Population als Hinweis darauf gesehen werden kann, dass die Stichprobe die anvisierte Grundgesamtheit gut repräsentiert (Wyschkon, 2011, S. 42).

Letztendlich bleibt somit fraglich, ob es prinzipiell überhaupt möglich ist, eine repräsentative Stichprobe zu untersuchen. Für die Normstichproben von LESEN 6-7 und LESEN 8-9 ist die Repräsentativität anzuzweifeln. Die zahlreichen Probleme im Zusammenhang mit der Repräsentativität von Normstichproben lassen die Möglichkeit der Definition von Niveaustufen, wie sie im Rahmen von IRT-basierten Tests möglich ist, attraktiv erscheinen. Sie ermöglichen eine stichprobenunabhängige Beurteilung des Leistungsniveaus. Aufgrund ihrer bereits in Kapitel 5.3.2.2 diskutierten gravierenden Nachteile (u. a. geringere Differenzierungsfähigkeit, keine Berücksichtigung des Messfehlers, künstliche Stufenbildung), wurde dennoch auf die Definition von Niveaustufen verzichtet.

Deskriptive Statistik und Verteilungen. Die Betrachtung der deskriptiven Statistiken und der Verteilungsformen zeigte zwar einerseits, dass in einigen Klassenstufen und Schularten die Verteilungen leicht schief ausfallen und zum Teil signifikant

von der Normalverteilung abweichen (wobei die signifikanten Abweichungen nicht zuletzt den zum Teil großen Stichprobenumfängen geschuldet sein dürften). Andererseits zeigten sich weder Boden- noch Deckeneffekte, und die Tests scheinen somit im gesamten Leistungsspektrum zu differenzieren, was als sehr zufriedenstellend bewertet wird.

Normtabellen. Auf Basis der Daten der Normstichproben wurden Tabellen mit Vergleichswerten erstellt. Für jede Klassenstufe existiert schulartübergreifend sowie nach Schularten getrennt eine Normtabelle, in der jedem Rohwert ein Prozentrang, ein T-Wert und ein T-Wertband zugeordnet ist. Weiter steht eine Tabelle zur Verfügung, in der jedem Rohwert ein Personenparameter mit entsprechendem Konfidenzintervall zugeordnet ist, sowie eine Tabelle mit kumulierten Häufigkeiten bezüglich der T-Wert-Differenz der Subtests. Insgesamt scheinen die Normtabellen und Interpretationshilfen in den Testmanualen von LESEN 6-7 und LESEN 8-9 ausreichend Informationen zur Verfügung zu stellen, um einerseits Ansatzpunkte für Fördermaßnahmen daraus ableiten zu können und andererseits die Daten für Forschungszwecke nutzen zu können.

Insgesamt kann also festgehalten werden, dass die Normstichproben von LESEN 6-7 und LESEN 8-9 ausreichend groß ausfallen und die Möglichkeiten zur Rekrutierung möglichst repräsentativer Stichproben weitgehend ausgeschöpft wurden, wenngleich Zweifel bezüglich der tatsächlichen Repräsentativität bleiben. Besonders positiv zu bewerten ist in Bezug auf LESEN 6-7 und LESEN 8-9, dass die Tests jeweils im gesamten Leistungsspektrum differenzieren, sowie dass in den Testmanualen ausführliche Informationen für verschiedene Nutzungsmöglichkeiten der Daten zur Verfügung gestellt werden.

Kapitel 15

Itemanalysen und Prüfung der Modellkonformität auf Basis der Normdaten

Auf Basis der Normdaten wurden noch einmal alle Items analysiert, um ihre Güte an einer großen Stichprobe zu überprüfen. Für den Subtest BLK wurden Hinweise auf die Angemessenheit der Itemschwierigkeit und der Zeitbegrenzung betrachtet. Für den Subtest TV erfolgte eine erneute Berechnung von KTT-Itemkennwerten sowie eine Einschätzung der Rasch-Modell-Konformität.

15.1 Methode

Im Folgenden wird zunächst das methodische Vorgehen für den Subtest BLK und anschließend jenes für den Subtest TV beider Tests dargestellt.

Subtest BLK. Da der Subtest BLK für beide Tests identisch ist, konnten die Analysen über die gesamte Normstichprobe von Klassenstufe sechs bis neun stattfinden ($N = 2\,584$). Um davon ausgehen zu können, dass die Items inhaltlich leicht genug sind, um aufgrund der Lesegeschwindigkeit und nicht aufgrund des Allgemeinwissens oder höherer Verständnisprozesse zu differenzieren, sollten nur wenige Fehler aufgetreten sein. Daher wurde überprüft, wieviel Prozent der Schüler die einzelnen Items korrekt lösten und ob Auslassungen und Fehler bei einzelnen Items gehäuft vorkamen. Außerdem wurde geprüft, ob Decken- oder Bodeneffekte vorliegen und ob insbesondere Itemanzahl, Itemschwierigkeit und Zeitbegrenzung angemessen gewählt wurden.

Subtest TV. Die Analysen bezüglich der Subtests TV erfolgten für beide Tests separat. Auf Basis der Normdaten wurden wie in den Voruntersuchungen Schwierigkeits- und Trennschärfeanalysen gemäß KTT durchgeführt. Es wurde zudem betrachtet, wie viel Prozent der Items von den Schülern der verschiedenen Klassenstufen und Schularten korrekt gelöst wurden. Weiter sollte eine deskriptive Analyse der Itemschwierigkeiten Aufschluss darüber geben, ob alle Schwierigkeitsbereiche abgedeckt sind, also ob es sowohl sehr schwere als auch sehr leichte Items gibt. Außerdem wurde für beide Tests jeweils die mittlere Schwierigkeit der Items beider Texte miteinander verglichen.

Im Rahmen der Trennschärfeanalysen wurden sodann sowohl die Trennschärfen aller Items auf Signifikanz geprüft als auch die mittlere Trennschärfe und Streubreite ermittelt. Zusätzlich erfolgte analog zu den Voruntersuchungen eine Berechnung der Selektionskennwerte (SK), um bei der Interpretation der Trennschärfewerte die Itemschwierigkeit mitzuberücksichtigen.

Ebenfalls wie in Kapitel 13.3.4 wurden zur Prüfung der Eindimensionalität Screetests, Parallelanalysen und MAP-Tests durchgeführt, zur Prüfung der lokalen Unabhängigkeit Q_3-Tests nach Yen (1984) angewendet und zur Prüfung des Item-Fit die wMNSQ-Werte herangezogen. Laut (Wilson, 2005, S. 129) ist bei „großen“ Stichproben zu erwarten, dass die *T*-Werte für einige Items signifikant ausfallen. Da die hier vorliegenden Stichproben mit $N = 1\,617$ für LESEN 6-7 und $N = 944$ für LESEN 8-9 als groß einzuordnen sind, wurden – wie von Wilson empfohlen – nur diejenigen Items als problematisch betrachtet, welche sowohl auffällige wMNSQ-Werte als auch *T*-Werte $> |2|$ aufweisen. Zur Prüfung der Personenhomogenität wurden wie in Kapitel 13.3.4 grafische Modelltests sowie Andersen-Tests durchgeführt. Als Teilungskriterium wurde hier wie zuvor jeweils der Median gewählt.

15.2 Ergebnisse

Auch die Ergebnisse werden nach Subtests aufgeteilt dargestellt. Zuerst wird auf die Ergebnisse bezüglich des Subtests BLK eingegangen und dann auf die Ergebnisse bezüglich des Subtests TV.

Subtest BLK. Beim Subtest BLK wurden alle Items im Mittel von 98.4 % der Schüler, die diese Items in Angriff nahmen, auch korrekt gelöst. Jedes einzelne Item wurde von mindestens 90 % der Schüler, die es in Angriff nahmen, korrekt gelöst. 49 der 100 Items wurden von mindestens 99 % der Schüler, die sie in Angriff nahmen, korrekt gelöst (s. Tab. 42 in Anhang C). 95 der 100 Items wurden immerhin noch von mindestens 95 % der Schüler, die sie in Angriff nahmen, korrekt gelöst. Die Fehler und Auslassungen verteilen sich gleichmäßig über alle Items, sodass nicht davon auszugehen ist, dass einzelne Items besonders schwer sind. Die Ergebnisse sprechen somit dafür, dass die Items ausreichend leicht sind, um basale Lesekompetenzen abzubilden. Insgesamt erreichte von den fast 2 600 Schülern nur ein Schüler die volle Punktzahl, d. h. er bearbeitete innerhalb von drei Minuten alle 100 Items korrekt. Die Betrachtung der Rohwertverteilungen (s. Abb. 18 in Kap. 14) zeigt ebenfalls, dass es keine Boden- und allenfalls leichte Deckeneffekte gibt.

Subtest TV. Im Subtest TV von LESEN 6-7 lösten die Schüler der sechsten Hauptschulklassen im Schnitt etwa 40 % der Aufgaben korrekt, während der Lösungsanteil bei den Schülern aus siebten Gymnasialklassen bei etwa 75 % liegt (s. Tab. 18). Bei LESEN 8-9 lösten die Hauptschüler der achten Klasse etwa 45 % der Aufgaben kor-

rekt, während die Schüler der neunten Gymnasialklassen im Schnitt etwa 75 % der Aufgaben korrekt lösten.

Tabelle 18. Anteil (in %) an Items, der von Schülern der verschiedenen Klassenstufen und Schularten korrekt gelöst wurde.

	LESEN 6-7					LESEN 8-9			
	HS	RS	GYM	GES		HS	RS	GYM	GES
6. Klasse	39	55	70	55	8. Klasse	44	53	70	55
7. Klasse	48	61	76	61	9. Klasse	47	60	74	62

Bei LESEN 6-7 fallen die *Schwierigkeitsindizes* für die Items zum expositorischen Text mit einer mittleren Schwierigkeit von $p = .54$ und einer Streubreite von $.24 < p < .77$ etwas niedriger aus als die Schwierigkeitsindizes für die Items zum narrativen Text mit einer mittleren Schwierigkeit $p = .63$ und einer Streubreite von $.44 < p < .86$. Bei LESEN 8-9 fallen die Schwierigkeitsindizes ebenfalls für die Items zum expositorischen Text etwas niedriger aus als die Werte für die Items zum narrativen Text, wobei der Unterschied hier kleiner ist. Die Items zum expositorischen Text haben eine mittlere Schwierigkeit von $p = .57$ mit einer Streubreite von $.23 < p < .86$, und die Items zum narrativen Text haben eine mittlere Schwierigkeit von $p = .60$ mit einer Streubreite von $.31 < p < .87$. Das Muster der Itemschwierigkeiten entspricht somit dem in Kapitel 13.3.4 ermittelten Muster der strukturellen Textschwierigkeit. Gemäß Flesch-Index und LIX sind die expositorischen Texte strukturell schwieriger als die narrativen Texte. Insgesamt liegen beim Subtest TV die Schwierigkeitsindizes der einzelnen Items für LESEN 6-7 zwischen $p = .24$ und $p = .86$ mit einem Mittelwert von $p = .59$. Bei LESEN 8-9 liegen die Werte zwischen $p = .23$ und $p = .87$ mit einem Mittelwert von $p = .58$. In beiden Tests sind also schwere und leichte Items enthalten, wobei im Mittel über die gesamte Normstichprobe jeweils knapp 60 % der Items richtig gelöst wurden. Die genauen Schwierigkeitsindizes für die einzelnen Items können Tabelle 43 in Anhang C entnommen werden.

Ebenfalls in Tabelle 43 in Anhang C befinden sich die Ergebnisse der *Trennschärfeanalysen*. Die Trennschärfeanalysen zeigen für LESEN 6-7, dass alle Trennschärfen signifikant von Null abweichen (alle $p < .01$). Die Werte der punktbiserialen Korrelationen liegen zwischen $r_{it} = .23$ und $r_{it} = .47$ und weisen somit eine Streubreite von $r_{it_{max}} - r_{it_{min}} = 0.24$ auf. Die mittlere Trennschärfe liegt bei $r_{it} = .38$. Die SK-Werte liegen zwischen .24 und .50. Somit sind alle SK $> .20$. Für LESEN 8-9 erwiesen sich ebenfalls alle Trennschärfen als signifikant von Null verschieden (alle $p < .01$). Die Trennschärfekoeffizienten weisen dabei eine Streuung von $r_{it} = .17$ bis $r_{it} = .50$ und somit eine Streubreite von 0.33 auf. Die mittlere Trennschärfe liegt bei $r_{it} = .36$. Die SK-Werte liegen zwischen .20 und .60. Zwei Items (10 und 16) weisen eine Trennschärfe von $r_{it} < .20$ auf, was als extrem niedrig zu bewerten ist. Der

SK liegt jedoch wie erwünscht für keines der Items unter SK $= .20$, was ein Hinweis darauf ist, dass die niedrigen Trennschärfen aufgrund sehr hoher oder sehr niedriger Itemschwierigkeit zustande kommen und somit vertretbar sind.

Zur Prüfung der *Eindimensionalität* wurden wie bereits in Kapitel 13.3.4 exploratorische Faktorenanalysen (EFAs) durchgeführt. Die für EFAs erforderliche Mindestgröße der Stichprobe von $N = 60$ ist für beide Tests gegeben. Die Stichprobe von LESEN 6-7 umfasst 1 617 Schüler, die Stichprobe von LESEN 8-9 umfasst 945 Schüler. Die Itemreliabilitäten sollten mindestens bei $r = .60$ liegen, um die Ladungen zuverlässig schätzen zu können (vgl. Bühner, 2011). Da die Stichproben hier groß sind ($N > 100$), sind auch etwas niedrigere Werte akzeptabel (vgl. MacCallum et al., 1999). Die Itemreliabilitäten werden, wie zuvor, über den jeweils höchsten Korrelationswert eines Items mit einem beliebigen anderen Item geschätzt. Bei LESEN 6-7 liegt nur für zwei Items der höchste Korrelationswert mit einem anderen Item über $r = .60$. Im Mittel beträgt der höchste Korrelationswert eines Items mit einem anderen Item $r = .40$, das Minimum liegt bei $r = .25$, das Maximum bei $r = .61$. Bei LESEN 8-9 liegt der höchste Korrelationswert eines Items mit einem anderen Item für kein Item über $r = .60$. Der Mittelwert liegt bei $r = .40$, das Minimum bei $r = .25$, das Maximum bei $r = .59$. Das bedeutet, dass die Korrelationsmatrizen nicht optimal für eine EFA geeignet sind. Allerdings wird das von Marcus und Bühner (2009) angegebene Mindestkriterium von $r = .20$ in keinem Test von einer Korrelation unterschritten.

Mithilfe des Kaiser-Meyer-Olkin- (KMO-) Koeffizienten wurde geprüft, ob substanzielle Korrelationen in der Korrelationsmatrix vorliegen. Bühner (2011) nennt hier für den KMO-Mindestwert von .50. Es ergeben sich mit KMO $= .94$ für LESEN 6-7 und KMO $= .92$ für LESEN 8-9 jeweils sehr gute Werte (vgl. Bühner, 2011). Zusätzlich wurde für jedes Item der Measure of Sample Adequacy- (MSA-) Koeffizient bestimmt. Dieser Wert sollte ebenfalls mindestens bei MSA $= .50$ liegen. Die MSA-Werte liegen bei LESEN 6-7 zwischen MSA $= .90$ und MSA $= .96$, was wiederum für eine sehr gute Eignung aller Items für eine EFA spricht. Bei LESEN 8-9 liegen die MSA-Werte zwischen MSA $= .78$ und MSA $= .95$, was für eine mittelgute bis sehr gute Eignung aller Items für eine EFA spricht. Schließlich wurde noch der Bartlett-Test durchgeführt. Er wird für beide Tests signifikant (beide $p < .01$), was bedeutet, dass alle Korrelationen größer als Null sind und die Matrix somit faktorisierbar ist. Bei Stichproben von $N > 60$ fällt der Bartlett-Test jedoch in den meisten Fällen signifikant aus. Ein signifikantes Ergebnis stellt hier daher nur ein Mindestkriterium dar (vgl. Bühner, 2011).

Insgesamt sprechen die meisten Kriterien für die Eignung der Normdaten für eine EFA. Lediglich bei der Reliabilitätenschätzung liegen die Werte für einige Items außerhalb des gewünschten Bereichs. Dabei handelt es sich jedoch nur um eine Mindestschätzung der Itemreliabilität und in keinem Test unterschreitet eine Korrelation den Mindestwert von $r = .20$. Da die Stichproben in beiden Fällen weit über $N = 100$ liegen, sind die Kriterien nicht so streng auszulegen, und eine EFA kann zumindest nützliche Hinweise über die Dimensionalität liefern.

Bei der Entscheidung über die Anzahl zu extrahierender Faktoren sprechen die Screeplots der auf der Basis tetrachorischer Korrelationen berechneten Eigenwerte bei beiden Tests für die Extraktion eines einzigen Faktors (s. Abb. 20; Eigenwerte s. auch Tab. 44 in Anhang C). Zum gleichen Ergebnis kommt jeweils der Minimum Average Partial- (MAP-) Test. Bei LESEN 6-7 ergibt sich die kleinste mittlere vierte Potenz der partiellen Korrelation bei einer auspartialisierten Komponente mit einem Wert von $< .0001$. Bei LESEN 8-9 fällt die mittlere vierte Potenz der partiellen Korrelation ebenfalls bei einer auspartialisierten Komponente am geringsten aus mit ebenfalls einem Wert von $< .0001$. Für Details siehe Tabelle 45 in Anhang C. Die Parallelanalyse spricht bei LESEN 6-7 für drei Faktoren, bei LESEN 8-9 für zwei Faktoren (s. Abb. 21). Bühner (2011) empfiehlt bei einem starken ersten Faktor den MAP-Test gegenüber der Parallelanalyse zu bevorzugen, da die Parallelanalyse im Rahmen einer Hauptachsenanalyse meist zu einer Überschätzung der Faktorenzahl kommt. Bei beiden Tests liegt ein starker erster Faktor vor, daher wird dem MAP-Test mehr Bedeutung zugeschrieben. Weiter empfiehlt Bühner stets beide Methoden (MAP-Test und Parallelanalyse) anzuwenden und bei unterschiedlichen Ergebnissen die besser interpretierbare zu wählen, was hier ebenfalls die Ein-Faktorenlösung ist. Laut Gorsuch (1983) und Kline (2002, S. 52) sollten Ladungen mindestens $\lambda = .30$ betragen. Bei LESEN 6-7 liegen bei der Ein-Faktorenlösung alle Ladungen über $\lambda = .30$, bei LESEN 8-9 liegen für zwei Items (10 und 16) die Ladungen darunter (s. Tab. 44 in Anhang C).

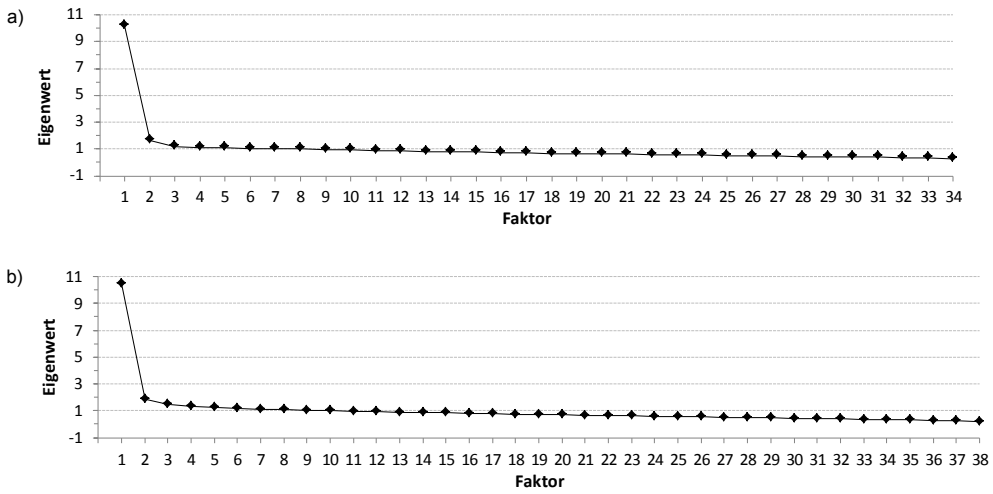


Abbildung 20. Eigenwert-Screeplots für LESEN 6-7 (a) und LESEN 8-9 (b).

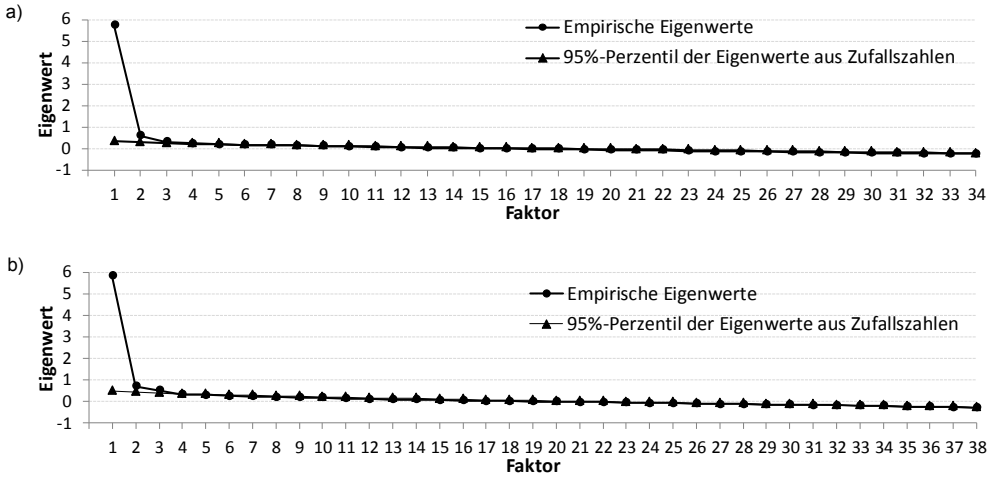


Abbildung 21. Ergebnisse der Parallelanalysen nach Horn für LESEN 6-7 (a) und LESEN 8-9 (b).

Die Auswertung bezüglich des *Item-Fits* ergab, dass bei LESEN 6-7 alle wMNSQ zwischen 0.94 und 1.15 und somit innerhalb des gewünschten Bereichs liegen. Bei zwölf der 34 Items ist $T > |2|$. Bei LESEN 8-9 liegen alle wMNSQ zwischen 0.86 und 1.15 und somit ebenfalls im gewünschten Bereich, wobei für 15 der 38 Items $T > |2|$ ist. Somit erfüllen alle Items des Subtests TV beider Tests mindestens eines der notwendigen Kriterien ($0.75 <$ wMNSQ < 1.33 oder $T \leq |2|$), was als zufriedenstellend gelten kann. Alle Ergebnisse der Itemanalysen finden sich in den Tabellen 46 und 47 in Anhang C.

Zur Prüfung der *Personenhomogenität* wurden erneut grafische Modelltests sowie Andersen-Tests durchgeführt. Beim grafischen Modelltest mit dem Median als Teilungskriterium zeigen sich bei LESEN 6-7 leichte und bei LESEN 8-9 etwas stärkere Abweichungen einiger Werte von der Winkelhalbierenden (s. Abb. 22). Zur Prüfung, ob die Abweichungen vom Modell signifikant sind, wurden daher in einem nächsten Schritt Andersen-Tests durchgeführt. Der Andersen-Test fällt sowohl bei LESEN 6-7 als auch bei LESEN 8-9 signifikant aus. Für LESEN 6-7 beträgt der Likelihood-Ratio-Wert 239.26 ($p < .01$; $\chi^2_{krit}(p = .80, df = 33) = 39.57$). Für LESEN 8-9 liegt der Likelihood-Ratio-Wert bei 254.83 ($p < .01$; $\chi^2_{krit}(p = .80, df = 37) = 43.98$).

Die Beträge der Q_3-Werte zur Beurteilung der *lokalen Unabhängigkeit* überschreiten weder bei LESEN 6-7 noch bei LESEN 8-9 die Grenze von $Q_3 = .20$. Für LESEN 6-7 liegen die Residualkorrelationen zwischen $Q_3 = .07$ und $Q_3 = .20$, für LESEN 8-9 liegen sie zwischen $Q_3 = .07$ und $Q_3 = .18$. Somit kann gemäß des Kriteriums von Yen (1984) für beide Tests lokale Unabhängigkeit der Items des Subtests TV angenommen werden.

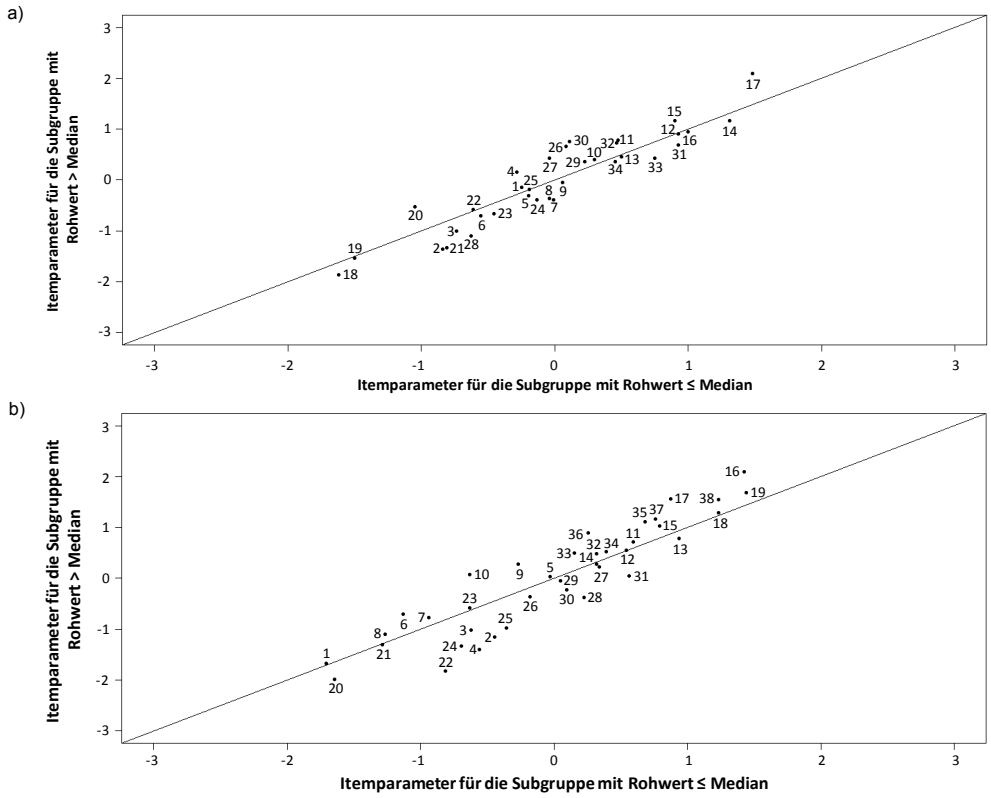


Abbildung 22. Grafischer Modelltest für LESEN 6-7 (a) und LESEN 8-9 (b); Teilungskriterium: Median des Summenwertes, Markierungsbeschriftung: Itemnummer.

15.3 Diskussion

Größtenteils bestätigen die Itemanalysen unter Verwendung der Normdaten die Ergebnisse der Voruntersuchungen und sprechen für eine hohe Itemgüte bei LESEN 6-7 und LESEN 8-9.

Subtest BLK. Beim Subtest BLK sprechen die Ergebnisse der Datenanalysen auf Item- und Personenebene dafür, dass die Items wie erwünscht sehr leicht sind. Nur sehr wenige Schüler kreuzten eine falsche Antwortalternative an. Es ist aber nicht auszuschließen, dass einige Schüler doch bei einzelnen Items länger überlegen mussten, ob diese nun richtig oder falsch sind, und daher aufgrund des Inhalts langsamer waren als sie es aufgrund der reinen Lesegeschwindigkeit gewesen wären. Insgesamt konnten nur wenige Schüler alle Items bearbeiten. Dies spricht dafür, dass die Be-

arbeitungszeit von drei Minuten knapp genug gewählt wurde, um Deckeneffekte zu vermeiden. Der Subtest BLK scheint im gesamten Leistungsspektrum von der sechsten Hauptschul- bis zur neunten Gymnasialklasse zwischen den Schülern differenzieren zu können.

Subtest TV. Die Ergebnisse der KTT-Itemanalysen zum Subtest TV sind sowohl hinsichtlich der Itemschwierigkeit als auch hinsichtlich der Trennschärfe zufriedenstellend. Die Selektionskennwerte aller Items liegen wie erwünscht über SK $= .20$. Beide Tests enthalten sowohl leichte als auch schwere Items und decken somit einen breiten Schwierigkeitsbereich ab, was eine Differenzierung im gesamten Leistungsspektrum erlaubt.

Die Ergebnisse der EFAs im Rahmen der dritten Voruntersuchung zum Subtest TV sprachen bereits für die Eindimensionalität des Subtests TV beider Tests. Bei den Analysen auf Basis der großen Normstichprobe spricht nun erneut Vieles für die Eindimensionalität. Neben den Ergebnissen der EFAs führten auch die Prüfung des Item-Fits und die Prüfung der lokalen Unabhängigkeit zu den erwünschten und erwarteten Ergebnissen, die ebenfalls die Eindimensionalitäts- und Homogenitätsannahme für die Subtests TV beider Tests stützen. Die Andersen-Modelltests auf Personenhomogenität fallen dagegen bei Verwendung des Medians als Teilungskriterium für beide Tests signifikant aus. Dieses Ergebnis spricht gegen ein Vorliegen von Personenhomogenität und somit dafür, dass nicht bei allen Personen eine homogene Fähigkeit gemessen wird bzw. dass die Personen der beiden gewählten Substichproben unterschiedliche Fähigkeiten zur Lösung der Items benutzten. Für LESEN 8-9 wurde somit das Ergebnis der dritten Voruntersuchung zum Subtest TV bestätigt, während für LESEN 6-7 das Ergebnis des Andersen-Tests bei der dritten Voruntersuchung zum Subtest TV nicht signifikant ausgefallen war. Zumindest im Fall von LESEN 6-7 könnte das Ergebnis hier lediglich wegen der großen Stichprobe signifikant ausfallen, obwohl die Abweichungen vom Modell gering sind. Der Andersen-Test gilt als sehr sensibel gegenüber Modellverletzungen (vgl. Bühner, 2011). Theoretisch hätte Personenhomogenität dadurch hergestellt werden können, dass Personen, die nicht modellkonform antworteten aus den Analysen ausgeschlossen werden. Dies wird bisweilen getan, ist jedoch sehr umstritten, da dadurch die Stichprobe verändert wird. Aus diesem Grund wurde in der vorliegenden Arbeit auf dieses Vorgehen verzichtet.

Ein Problem im Zusammenhang mit der Personenhomogenität könnte bei LESEN 6-7 und LESEN 8-9 wie bereits erläutert aus der Möglichkeit resultieren, bei einem SC-Antwortformat auch nach dem Ausschlussverfahren vorgehen oder raten zu können, sowie daraus, dass Schüler bei Nichtwissen unterschiedliche Strategien anwenden können (Item auslassen oder raten). Eine weitere Erklärungsmöglichkeit könnte die Zeitbegrenzung darstellen, denn diese ermöglicht es ebenfalls, zwei unterschiedliche Strategien anzuwenden: Entweder eine erste Strategie, die Aufgaben besonders genau und fehlerfrei zu lösen, oder eine zweite Strategie, möglichst alle Aufgaben in der vor-

gegeben Zeit zu bearbeiten. Zwar wurde die Zeitbegrenzung so großzügig gewählt, dass alle Schüler alle Aufgaben bearbeiten können sollten, dennoch ist nicht ganz auszuschließen, dass auch die zweite Strategie angewendet wird. Wilhelm und Schulze (2002) zeigten im Zusammenhang mit Intelligenztests, dass eine Zeitbegrenzung bei Power-Tests nicht zu vernachlässigen ist. Zeitbegrenzte Intelligenztests korrelierten deutlich höher mit zeitbegrenzten Speedtests als mit Intelligenztests ohne Zeitbegrenzung. Die Autoren schlussfolgerten daraus, dass zeitbegrenzte Intelligenztests noch etwas anderes als Intelligenz erfassen, z. B. Verarbeitungsgeschwindigkeit oder Sorgfalt der Aufgabenbearbeitung. Dies könnte auch bei LESEN 6-7 und LESEN 8-9 der Fall sein. Jedoch erscheint es nicht praktikabel, einen Test völlig ohne Zeitbegrenzung vorzugeben (vgl. J. Rost, 2004, S. 43). Gymnasiasten aller Klassenstufen brauchten zum größten Teil nur etwa die Hälfte der maximal verfügbaren Zeit. Daher und um die Durchführungsdauer nicht über eine Schulstunde hinaus auszudehnen, wurde die maximal zur Verfügung stehende Zeit nicht erhöht oder gar ganz auf eine Zeitbegrenzung verzichtet.

Bei der Interpretation der Ergebnisse der Andersen-Tests ist außerdem zu berücksichtigen, dass derartige Modelltests generell umstritten sind. Die Festlegung, ab wann eine Abweichung vom Modell signifikant ist, ist stets willkürlich bzw. auf reiner Konvention basierend. Meist besteht ein Dilemma zwischen der Güte der Modellpassung und dem Aufwand, mit dem diese erreicht wird – also der Frage nach der Gewichtung des Kriteriums der empirischen Gültigkeit im Verhältnis zum Einfachheitskriterium. Für die Stärke der Abweichungen vom Rasch-Modell stehen leider keine Effektgrößen zur Verfügung, die es erlauben würden, eine angemessene Stichprobengröße für einen Modelltest mit optimaler Teststärke und Irrtumswahrscheinlichkeit α zu bestimmen (vgl. Bühner, 2011). Daher ist das Ergebnis stets abhängig von der Stichprobengröße, und bei großen Stichproben – wie den Normstichproben von LESEN 6-7 und LESEN 8-9 – ist ein Erreichen des Signifikanzniveaus selbst bei minimalen Abweichungen vom Modell wahrscheinlich. Einige Autoren (z. B. K. D. Kubinger, 2006; Bühner, 2011; Strobl, 2010) halten Modelltests dennoch für unabdingbar und empfehlen deren Anwendung, während andere Autoren die Tests kritisch bewerten oder gar ganz darauf verzichten und sich lediglich auf die Prüfung des Item-Fits beschränken (vgl. z. B. DeMars, 2010, S. 57ff.).

Die Ergebnisse der vorgängig beschriebenen Untersuchungen sprechen insgesamt für eine hohe Qualität der Items. Bis auf die umstrittenen Andersen-Tests sprechen zudem alle Analysen für Rasch-Modell-Konformität und stützen die Eindimensionalitätsannahme für beide Tests. Somit spricht Vieles dafür, dass die laut Verrechnungsregel resultierenden Ergebniswerte die empirischen Merkmalsrelationen adäquat abbilden, wodurch das Gütekriterium der Skalierbarkeit gegeben scheint (vgl. Moosbrugger & Kelava, 2008, S. 18). Dies ist eine Voraussetzung für Reliabilität und Validität der Tests. Auf diese beiden Gütekriterien hin werden LESEN 6-7 und LESEN 8-9 in den folgenden Kapiteln eingehend geprüft.

Kapitel 16

Reliabilitätsanalysen

Im Rahmen der Reliabilitätsanalysen wurde mithilfe verschiedener Methoden die Zuverlässigkeit von LESEN 6-7 und LESEN 8-9 überprüft. Hierfür wurden zum einen die üblichen KTT-Ansätze verwendet, nämlich verschiedene Konsistenzmethoden sowie die Retestmethode (eine Paralleltestkorrelation wurde nicht berechnet, da keine parallelen Tests verfügbar sind). Zum anderen wurden IRT-Ansätze zur Beurteilung der Reliabilität herangezogen, nämlich die Erwartungswertmethode sowie der EAP/PV-Schätzer.

16.1 Methode

Die Koeffizienten der Konsistenz- und IRT-Methoden wurden auf Basis der Normdaten berechnet (Beschreibung der Normstichprobe s. Kap. 14.2). Für die Retestmethode fand eine separate Datenerhebung statt, die im entsprechenden Abschnitt beschrieben wird.

Konsistenzmethoden. Zur Bestimmung der Konsistenz wurde zum einen die interne Konsistenz berechnet und zum anderen wurden verschiedene Testhalbierungsmethoden angewendet. Die *interne Konsistenz* liefert nur bei homogenen Tests einen eindeutig interpretierbaren Koeffizienten (Lienert & Raatz, 1998, S. 200). Sie wurde für beide Subtests bestimmt. Wie zuvor im Rahmen der Testkonstruktion (s. Kap. 13.3) wurde auch hier die KR-20-Formel für dichotome Daten angewendet (Lienert & Raatz, 1998, S. 193). Da beim Subtest BLK die Items ausschließlich aufgrund ihrer Homogenität ausgewählt wurden, sollten hohe Konsistenzwerte hier nicht überbewertet werden. Für den Gesamttest wurde keine interne Konsistenz ermittelt, da es keinen Sinn macht, Speedtest-Items mit Niveautest-Items zu vermischen, und es sich um eine inhomogene Skala handeln würde.

Für die *Testhalbierung* gibt es verschiedene Möglichkeiten (s. Kap. 5.3.2.1). In der vorliegenden Arbeit wurden die „Odd-even-Testhalbierungsmethode“ und die Methode der Bildung von „Itemzwillingen“ gewählt. Da, wie erläutert, beim Subtest BLK Reliabilitätswerte bezüglich der Konsistenz nicht sehr aussagekräftig sind, werden diese aufwendiger durchzuführenden Testhalbierungsmethoden nur beim Subtest TV angewendet. Bei der Odd-even-Testhalbierungsmethode wurden für jeden Subtest die

Items mit gerader Reihungsnummer mit denjenigen Items mit ungerader Reihungsnummer korreliert und die Korrelation mithilfe einer Korrekturformel nach Spearman-Brown aufgewertet (Bühner, 2011, S. 157, 162). Die Aufwertung erfolgte, da nur „halbe“ Tests korreliert werden, obwohl der komplette Test durchgeführt wurde. Die Odd-even-Methode wurde gewählt, damit sich Übungs- und Ermüdungseffekte gleichmäßig über beide Testhälften verteilen. Darüber hinaus wurde eine Testhalbierung auf Basis der Analysedaten durchgeführt (Lienert & Raatz, 1998, S. 183). D. h. es wurden sogenannte „Itemzwillinge“ mit ähnlicher Schwierigkeit und Trennschärfe gebildet, die Zwillingspaare jeweils auf zwei Gruppen aufgeteilt und anschließend die Ergebniswerte beider Gruppen korreliert.

Retestmethode. Da Konsistenzwerte für den Subtest BLK wenig aussagekräftig sind, wurde zur weiteren Reliabilitätsprüfung zusätzlich die Retestmethode angewendet. Bei der Retestmethode fließen in den Reliabilitätskoeffizienten neben der Reliabilität des Tests auch immer sowohl die Stabilität des Merkmals als auch die Konstanz der äußeren Bedingungen mit ein (Lienert & Raatz, 1998, S. 201). Um die Retestkoeffizienten zu bestimmen, wurden beide Tests bei einer kleineren, ausschließlich hierfür rekrutierten Stichprobe mit einem Abstand von drei Wochen wiederholt durchgeführt. Die Ergebnisse der Subtests und des Gesamtergebnisses des ersten Messzeitpunktes wurden mit den entsprechenden Ergebnissen des zweiten Messzeitpunktes korreliert. Der daraus resultierende Retestkoeffizient kann als weiterer Schätzer für die Reliabilität angesehen werden. Für das Gesamtergebnis erfolgte die Korrelation auf Klassenstufenebene, da die Normwerte klassenstufenspezifisch berechnet wurden. Die Subtest-T-Werte zur Ermittlung des Gesamtergebnisses wurden den Normtabellen mit den schulartübergreifenden T-Werten entnommen.

Aufgrund der großen Itemanzahl – vor allem im Subtest BLK – und dem Fehlen einer Rückmeldung über die Korrektheit der Item-Beantwortung wurde davon ausgegangen, dass kaum Lerneffekte auftreten und ein zeitliches Intervall von drei Wochen zwischen den Messungen ausreicht. Zur Absicherung wurde das Auftreten von Wiederholungseffekten anschließend überprüft. Hierfür wurde für beide Subtests und das Gesamtergebnis neben einer Prüfung auf signifikante Unterschiede zwischen den Messzeitpunkten anhand eines t-Tests für abhängige Stichproben auch die *Effektstärke* (d) ermittelt. Die Effektstärke stellt die standardisierte Mittelwertsdifferenz dar (Cohen, 1988, S. 20). Effekte von $d < 0.2$ gelten als vernachlässigbar, Effekte ab $d = 0.2$ als klein, Effekte ab $d = 0.5$ als moderat und Effekte ab $d = 0.8$ als groß (Cohen, 1988, S. 24ff.). Auf eine Normalverteilungsprüfung, welche bei kleinen Stichproben eine Voraussetzung für den t-Test ist, konnte verzichtet werden, da alle Stichprobenumfänge $n \geq 30$ betrugen und der t-Test in diesem Fall robust gegenüber einer Verletzung der Normalverteilungsvoraussetzung reagiert (Bortz & Schuster, 2010; K. Kubinger, Rasch & Moder, 2009). Darüber hinaus wurden *Intraklassenkorrelationen* (ICC; Zwei-Weg, gemischt, Einzelwert, absolute Übereinstimmung) zwischen den Ergebnissen beider

Messzeitpunkte berechnet, um nicht nur die relative, sondern auch die absolute Stabilität der Ergebnisse zu berücksichtigen (Bühner, 2011).

Die Messwiederholung wurde im Februar 2011 durchgeführt und beschränkte sich aus praktisch-organisatorischen und ökonomischen Gründen auf den süddeutschen Raum (Bayern und Baden-Württemberg). Es kamen die Endversionen von LESEN 6-7 und LESEN 8-9 zum Einsatz. Weiter wurden die gleichen Zusatzinformationen wie bei der Normierung erfragt (Klassenstufe, Schulart, Geschlecht, Muttersprache, LRS-Diagnose, Deutschnote im letzten Zeugnis, Geburtsmonat und -jahr). Für LESEN 6-7 umfasste die Stichprobe 284 Schüler und für LESEN 8-9 waren es 240 Schüler aus jeweils ein bis zwei Klassen der Klassenstufen sechs bis neun und der Schularten Hauptschule, Realschule und Gymnasium sowie einzelne M-Zug-Schüler. Die Verteilung der Schüler über die Klassenstufen und Schularten kann Tabelle 19 entnommen werden.

Tabelle 19. Stichproben der Retestmessung.

LESEN 6-7						LESEN 8-9					
	HS	RS	GYM	AN	Σ		HS	RS	GYM	AN	Σ
6. Klasse	48	43	50	0	141	8. Klasse	35	40	53	25	153
7. Klasse	35	33	53	34	143	9. Klasse	19	12	41	15	87
Σ	83	76	103	22	284	Σ	54	52	94	40	240

Da es sich bei der Retestmessung um eine kleinere und weniger repräsentative Stichprobe handelt, wurde wie von Lienert und Raatz (1998, S. 204ff.) empfohlen überprüft, ob die Streuung der Werte in der kleineren Stichprobe derjenigen der Normstichprobe (bzw. eigentlich der Population) entspricht. Eine kleinere Streuung ist bei einer kleineren und weniger repräsentativen Stichprobe wahrscheinlich und wirkt sich negativ auf die Höhe der Reliabilitätskoeffizienten aus. Die tatsächliche Korrelationshöhe wird dadurch also unterschätzt. Daher wurde die von Lienert und Raatz (1998, S. 206) für diesen Fall empfohlene Korrekturformel zur Berücksichtigung der Streuungsunterschiede bei der Bestimmung des Retestkoeffizienten angewendet[9].

IRT-Methoden. Der Vorteil der IRT-Methoden zur Reliabilitätsschätzung liegt darin, dass diese die Fehlervarianz und/oder die wahre Varianz direkt schätzen können, ohne wie die KTT-Methoden einen Umweg über Korrelationen machen zu müssen (J. Rost, 2004, S. 380; vgl. Kap. 5.3.2.1). Die *Erwartungswertmethode* setzt den gemittelten Standardfehler der Personenparameterschätzung zur Varianz der Personenparameter ins Verhältnis. Der Index wurde wie von von J. Rost (2004, S. 380) beschrieben bestimmt. Der *EAP/PV-Reliabilitätsindex* wurde mithilfe des Programms ConQuest berechnet. Er setzt die individuellen erwarteten a posteriori Schätzwerte ins Verhältnis zur geschätzten Gesamtvarianz der latenten Fähigkeit.

[9] Abweichungen der Retestreliabilitätswerte von den in den Testmanualen angegebenen Werten sind darauf zurückzuführen, dass die Werte im Manual ohne diese Korrektur berechnet wurden.

16.2 Ergebnisse

Im Folgenden werden zunächst die Ergebnisse der Konsistenzmethoden berichtet, bevor auf die Ergebnisse der Retest-Analysen und schließlich auf die anhand der IRT-Methoden ermittelten Reliabilitätswerte eingegangen wird.

Konsistenzmethoden. Die interne Konsistenz für den Subtest BLK fällt sowohl für LESEN 6-7 ($N = 1\,585$) als auch für LESEN 8-9 ($N = 936$) mit einem Wert von jeweils KR-20 $= .97$ erwartungskonform sehr hoch aus. Beim Subtest TV liegen für LESEN 6-7 ($N = 1\,557$) alle berechneten Konsistenzkoeffizienten (interne Konsistenz, Odd-Even-Koeffizient und Item-Zwillings-Koeffizient) bei KR-20 $= r_{tt} = .87$. Für LESEN 8-9 liegt der Wert für die interne Konsistenz bei KR-20 $= .86$, der Odd-Even-Koeffizient bei $r_{tt} = .88$ und der auf der Basis von Itemzwillingen berechnete Koeffizient bei $r_{tt} = .87$. Somit fallen alle Konsistenzwerte sehr zufriedenstellend aus.

Retestmethode. Im Rahmen der Retest-Analysen wurde zunächst für beide Messzeitpunkte eine deskriptive Statistik berechnet (s. Tab. 20). Bei einer Betrachtung der Mittelwerte wird deutlich, dass diese bei beiden Tests für beide Subtests beim späteren, zweiten Messzeitpunkt höher liegen als beim früheren, ersten Messzeitpunkt. Ein Vergleich der Streuung der Retestdaten (s. Tab. 20) mit der Streuung der Normdaten (s. Tab. 17) zeigt zudem, dass die Normdaten eine größere Streuung aufweisen. Aus diesem Grund wurde bei der Bestimmung der Retest-Koeffizienten die von Lienert und Raatz (1998, S. 206) für diesen Fall empfohlene Korrekturformel angewendet.

Tabelle 20. Deskriptive Statistik für die Messwiederholungsdaten zum ersten Messzeitpunkt (t1) und zum zweiten Messzeitpunkt (t2).

LESEN 6-7				LESEN 8-9			
	N	*M*	*SD*		*N*	*M*	*SD*
BLK (t1)	283	53.70	12.44	BLK (t1)	236	62.57	12.39
BLK (t2)	280	61.25	13.81	BLK (t2)	226	69.85	13.80
TV (t1)	262	20.07	6.70	TV (t1)	211	21.96	6.13
TV (t2)	280	21.06	6.59	TV (t2)	226	24.27	6.79
GES (t1, 6. Klasse)	141	98.35	15.37	GES (t1, 8. Klasse)	149	103.46	14.31
GES (t2, 6. Klasse)	138	105.91	15.41	GES (t2, 8. Klasse)	139	110.63	14.98
GES (t1, 7. Klasse)	121	103.64	13.73	GES (t1, 9. Klasse)	62	105.15	11.38
GES (t2, 7. Klasse)	142	108.01	14.23	GES (t1, 9. Klasse)	84	110.35	13.47

Alle Retest-Koeffizienten (r_{tt}) fallen schlussendlich sehr zufriedenstellend aus (s. Tab. 21). Alle Korrelationen sind als signifikant einzustufen. Ebenfalls als signifikant einzustufen sind allerdings alle anhand von *t*-Tests für abhängige Stichproben geprüften Unterschiede zwischen dem ersten und zweiten Messzeitpunkt (alle $p < .01$).

Wider Erwarten zeigen sich vor allem beim Subtest BLK beider Tests sowie beim Gesamtergebnis in der sechsten Klassenstufe Wiederholungseffekte mittlerer Stärke, während die Effekte für die übrigen Klassenstufen und Subtests als recht klein bis unbedeutend zu bezeichnen sind. Dieses Muster spiegelt sich auch in den ICC-Koeffizienten wider, die für den Subtest BLK deutlich niedriger ausfallen und infolgedessen auch für das Gesamtergebnis relativ niedrig sind.

Tabelle 21. Ergebnisse der Retestanalysen.

LESEN 6-7					LESEN 8-9				
	N	r_{tt}	ICC	d		N	r_{tt}	ICC	d
BLK	278	.83	.66	0.57	BLK	222	.86	.67	0.56
TV	257	.87	.84	0.15	TV	197	.88	.82	0.36
GES 6. Klasse	138	.91	.77	0.52	GES 8. Klasse	149	.86	.74	0.49
GES 7. Klasse	121	.92	.72	0.29	GES 9. Klasse	58	.80	.65	0.42

Im Vergleich mit den bisher genannten KTT-Ergebnissen kommen die IRT-Methoden für die Subtests TV zu ähnlichen bis leicht niedrigeren Reliabilitätswerten. Der EAP/PV-Koeffizient beträgt für beide Tests EAP/PV $= .86$, die Erwartungswertmethode ergibt für beide Tests einen Wert von .85. Erwartungsgemäß stimmen die EAP/PV-Koeffizienten gut mit den Werten für die interne Konsistenz überein. Dass die Ergebnisse der Erwartungswertmethode etwas niedriger ausfallen war ebenfalls zu erwarten, da die Erwartungswertmethode die Reliabilität tendenziell eher unterschätzt (vgl. J. Rost, 2004, S. 382).

Tabelle 48 in Anhang D enthält alle Reliabilitätswerte für die einzelnen Klassenstufen sowie nach Klassenstufe und Schulart aufgeteilt. Vor allem bei Substichproben mit sehr kleiner Fallzahl fallen die Werte zum Teil niedriger aus, insbesondere der Retest-Reliabilitätswert für die neunte Hauptschulklasse ist sehr niedrig. Hier liegen allerdings nur Daten von 17 Schülern vor.

16.3 Diskussion

Im Rahmen der Reliabilitätsanalysen wurden verschiedene Reliabilitätsindizes betrachtet. Alle berechneten Konsistenzmaße sprechen für eine mittelhohe bis hohe Reliabilität beider Subtests von LESEN 6-7 und LESEN 8-9 und sind somit sehr zufriedenstellend. Da für den Subtest BLK Konsistenzmethoden wenig aussagekräftig sind, wurde eine zusätzliche Datenerhebung mit einer Messwiederholung nach drei Wochen durchgeführt, um Retestkoeffizienten berechnen zu können. Die Retestkoeffizienten fallen alle in den Bereich mittlerer Höhe und können somit ebenfalls als zufriedenstellend gelten. Allerdings zeigten sich erwartungswidrig Wiederholungseffekte, die zwar

größtenteils niedrig, jedoch beim Subtest BLK und beim Gesamtergebnis der sechsten Klasse mittelhoch ausfallen.

Signifikante Unterschiede zwischen den Messzeitpunkten der Retestmessung können auf verschiedene Ursachen zurückzuführen sein: Erstens könnten z. B. Übungseffekte dafür verantwortlich sein. In diesem Fall wäre möglicherweise ein größeres Retestintervall sinnvoll gewesen. Beim Subtest BLK könnte das Aufgabenformat für die Schüler bei der ersten Testung zunächst noch ungewohnt gewesen sein, bei der zweiten Testung nicht mehr. Beim Subtest TV kann jedoch davon ausgegangen werden, dass das Beantworten von Fragen zu einem Text im SC-Format den Schülern geläufig war. Es ist natürlich auch nicht auszuschließen, dass die Schüler sich nach der Testung über Items austauschten oder die Lösung nachlasen. Eine zweite Erklärungsmöglichkeit wäre, dass es sich beim Leseverständnis um ein instabiles Merkmal handelt bzw. die Schüler innerhalb der drei Wochen entsprechende Fortschritte im Leseverständnis gemacht haben. Dies ist jedoch unwahrscheinlich, wenn man bedenkt, wie gering die Leistungszuwächse in der Sekundarstufe in der Regel über ein ganzes Schuljahr hinweg sind. Eine dritte Erklärungsmöglichkeit wäre, dass LESEN 6-7 und LESEN 8-9 ungenau messen und so große Unterschiede zwischen den Messungen zustande kommen. Da sich die Leistungen der Schüler vom ersten zum zweiten Messzeitpunkt fast durchweg verbesserten und alle Reliabilitätsmaße (einschließlich der Retestreliabilität) auf eine zufriedenstellende bis sehr hohe Reliabilität hinweisen, erscheint diese Erklärung relativ unwahrscheinlich. Am wahrscheinlichsten erscheinen Übungs- und Lerneffekte als Ursache für die Leistungssteigerung.

Insgesamt können alle ermittelten Reliabilitätswerte für beide Subtests beider Tests als zufriedenstellend bis sehr gut beurteilt werden. Beide Tests scheinen sowohl für die Beurteilung von Gruppendifferenzen und interindividuellen Differenzen als auch für die Individualdiagnostik ausreichend reliabel zu sein. Zwar hätte eine weitere sukzessive Entfernung von Items mit relativ geringer Trennschärfe zu einer weiteren Verbesserung einiger Reliabilitätswerte führen können, allerdings warnt z. B. Bühner (2011, S. 256) davor, einen Test „zu Tode“ zu homogenisieren, da dies stets zu Lasten der Validität geschieht. Mit der Validität von LESEN 6-7 und LESEN 8-9 beschäftigt sich das nächste Kapitel.

Kapitel 17

Validitätsanalysen

Ziel der Validitätsanalysen war es, zu prüfen, ob LESEN 6-7 und LESEN 8-9 auch tatsächlich das Merkmal messen, das sie messen sollen, nämlich Leseverständnis. Im Folgenden werden die drei gängigsten Validitätsaspekte betrachtet, nämlich die Inhaltsvalidität, die Konstruktvalidität und die Kriteriumsvalidität (vgl. Kap. 5.3.2.1).

17.1 Generelles zur Methodik

Zur Beurteilung der inhaltlichen Validität werden in erster Linie theoretische Überlegungen angestellt und die Plausibilität der Operationalisierungen wird erörtert. Zur Beurteilung der Konstruktvalidität und der Kriteriumsvalidität hingegen werden statistische Kennwerte berichtet.

Um die Konstruktvalidität zu bestimmen, wurden zum einen die Testergebnisse von LESEN 6-7 und LESEN 8-9 mit den Ergebnissen anderer Lesetests korreliert. Zum anderen wurde geprüft, ob LESEN 6-7 und LESEN 8-9 zu verschiedenen aus den im ersten Teil der vorliegenden Arbeit beschriebenen theoretischen Überlegungen und empirischen Befunden abgeleiteten hypothetischen Ergebnissen kommen.

Zur Bestimmung der Kriteriumsvalidität wurden mehrere Außenkriterien erhoben und mit den Ergebnissen von LESEN 6-7 und LESEN 8-9 korreliert und damit Annahmen im Hinblick auf konvergente und diskriminante Validität geprüft. Zudem wurde untersucht, ob sich empirische Befunde zum Zusammenhang von Leseverständnis mit verschiedenen weiteren Außenkriterien anhand von LESEN 6-7 und LESEN 8-9 replizieren lassen.

Um die entsprechenden statistischen Kennwerte berechnen zu können, war die Erhebung weiterer Daten erforderlich. Die erneute und erweiterte Datenerhebung sowie die herangezogenen Stichproben werden im Folgenden beschrieben.

Datenerhebung und Stichproben. Die Datenerhebung für die Validitätsanalysen fand zu Beginn des Schuljahres 2010/2011 statt. Es wurden Hauptschul-, Realschul- und Gymnasialklassen der Klassenstufen sieben bis zehn rekrutiert. Die Testungen fanden, wie bereits in Kapitel 14 beschrieben, zu Beginn eines neuen Schuljahres und darum jeweils eine Klassenstufe höher statt. Entsprechend wurde in den Klassenstufen sieben und acht LESEN 6-7 durchgeführt und in den Klassenstufen neun und zehn LESEN 8-9. Um Irritationen zu vermeiden, werden auch in diesem Kapitel die Daten

im Folgenden den Klassenstufen zugeordnet, für die sie in der Normierung verwendet wurden (die Daten der siebten Klasse werden demnach unter Klassenstufe sechs geführt, die Daten der achten Klasse unter Klassenstufe sieben etc.).

Aus praktisch-organisatorischen und ökonomischen Gründen beschränkte sich die Datenerhebung auf den süddeutschen Raum. Daten für die neunte Klassenstufe der Hauptschule liegen nicht vor, da die Hauptschule in Bayern und Baden-Württemberg mit der neunten Klassenstufe abschließt und somit keine Erhebung in der nächsthöheren Klassenstufe stattfinden konnte. Insgesamt nahmen 452 Sechst- und Siebtklässler sowie 275 Acht- und Neuntklässler an den Datenerhebungen zur Validierung teil. Tabelle 22 zeigt die Verteilung der Stichprobe über die Schularten und Klassenstufen.

Tabelle 22. Schülerzahlen der Validierungsstichprobe pro Klassenstufe und Schulart.

	HS	RS	GYM	AN	Σ
6. Klasse	81	53	82	21	237
7. Klasse	85	49	57	24	215
8. Klasse	40	54	70	6	170
9. Klasse	—	53	51	—	104

Erhebungsinstrumente und Codierung. Zur Validierung wurden für die Schüler ein umfangreicher Schülerbogen und für die Deutschlehrkräfte ein Lehrerbogen erstellt. Der *Schülerbogen* umfasste neben LESEN 6-7 bzw. LESEN 8-9 und den Zusatzinformationen, die bereits bei der Normierung und Retestmessung erfragt wurden (Klassenstufe, Schulart, Geschlecht, Geburtsmonat und -jahr, Vorliegen einer LRS-Diagnose und Deutschnote im letzten Zeugnis), weitere Lesetests zur Konstruktvalidierung. Dabei handelte es sich um die in Kapitel 5.3.4 beschriebenen Tests LGVT 6-12 (Schneider et al., 2007) zur Erfassung von Lesegeschwindigkeit und Leseverständnis sowie WLST 7-12 (Schlagmüller & Schneider, 2007) zur Erfassung des Wissens über Lesestrategien. Zur Erfassung des Lesestrategiewissens in der sechsten Klassenstufe wurde auf einen für die sechste Klassenstufe entwickelten Metakognitionstest von Lingel et al. (2010) zurückgegriffen (im Folgenden mit „MKT“ abgekürzt). Weiter wurden Informationen zum familiären Hintergrund (höchster Schulabschluss beider Elternteile, Berufe beider Elternteile, Anzahl der Bücher zu Hause) und zum eigenen Leseverhalten („Wie oft liest du zu deinem Vergnügen?“) erfragt. Insgesamt dauerten die Testungen jeweils zwei Schulstunden.

Die angegebenen Berufe der Eltern wurden gemäß dem Bundesinstitut für Berufsbildung (2010, S. 2) klassifiziert und folgendermaßen codiert: 0 = Arbeitslos oder Hausfrau, 1 = Helfer, einfacher Dienst, 2 = Fachkräfte, mittlerer Dienst, 3 = Meister, Techniker, kaufmännische Fortbildungsberufe, Bachelorberufe, gehobener Dienst und 4 = (mind. 4-jährige) Studienberufe, höherer Dienst sowie Aufsichts- und Führungskräfte. Die Schulabschlüsse der Eltern wurden mit 0 = kein Schulabschluss, 1 = Hauptschulabschluss, 2 = Realschulabschluss und 3 = Abitur kategorisiert. Wie bereits in anderen

Studien (z. B. McElvany, Becker & Lüdtke, 2009; Wendt, Grölich, Guill, Scharenberg & Boss, 2010) wurde als Indikator des SÖS bei Elternteilen mit unterschiedlicher Berufskategorie der jeweils höhere Statuswert herangezogen. Bei Eltern mit unterschiedlichem Schulabschluss wurde ebenfalls jeweils der höhere Schulabschluss zur Bestimmung des Bildungshintergrunds gewählt. Das Item „Wie viele Bücher gibt es bei dir zu Hause?“ gab zusätzlich den Hinweis, dass auf einen Meter Bücherregal etwa 40 Bücher passen, und wurde mit folgender Antwortskala dargeboten: „0 bis 10“, „11 bis 25“, „26 bis 100“, „101 bis 200“, „201 bis 500“ oder „Mehr als 500“. Die Codierung erfolgte in dieser Reihenfolge aufsteigend von 0 bis 5. Die Antwortskala für das Item „Wie oft liest du zu deinem Vergnügen?“ lautete: „Nie oder fast nie“, „Ein paar Mal im Jahr“, „Etwa einmal im Monat“, „Etwa einmal pro Woche“ oder „Täglich oder fast täglich“. Die Codierung erfolgte in dieser Reihenfolge aufsteigend von 0 bis 4.

Im *Lehrerbogen* wurden als Außenkriterien zur Feststellung der Kriteriumsvalidität für jeden Schüler die Deutsch- und Mathematiknote sowie der Gesamtnotenschnitt im letzten Schulzeugnis erfragt. Zusätzlich sollte die Deutschlehrkraft auf einer zehnstufigen Skala von 1 (= sehr schlecht) bis 10 (= sehr gut) ein Urteil über die Lesekompetenz jedes einzelnen Schülers abgeben. Es wurde bewusst keine sechsstufige Notenskala verwendet, da diese – obwohl sie den Lehrkräften am besten vertraut sein dürfte – vermutlich dazu verleitet hätte, schlicht die Deutschnote anzugeben.

Auswertungen zum Gesamtergebnis. Da der Rohwert des Gesamtergebnisses für LESEN 6-7 und LESEN 8-9 aus der Summe der auf Klassenstufenebene ermittelten T-Werte der Subtests resultiert, ist es nicht sinnvoll, das Gesamtergebnis in klassenstufenübergreifende Berechnungen einzubeziehen. Der Mittelwert des Gesamtergebnisses ist für alle Klassenstufen 100. Somit können Schüler aus unterschiedlichen Klassenstufen dasselbe Gesamtergebnis als Rohwert erzielen, obwohl sie eigentlich ganz unterschiedliche Leistungen erbracht haben. Daher werden Berechnungen für das Gesamtergebnis nur klassenstufenweise durchgeführt; ein Klassenstufenvergleich erfolgt in Bezug auf das Gesamtergebnis nicht.

17.2 Inhaltsvalidität

Die Inhaltsvalidität lässt sich nicht statistisch überprüfen. Es existieren nicht wie bei der Reliabilität oder bei den anderen Validitätsaspekten Kennwerte, die typischerweise berechnet werden. Vielmehr wird die Inhaltsvalidität in der Regel aufgrund theoretischer, fachlich-logischer Überlegungen eingeschätzt (vgl. Bühner, 2011). LESEN 6-7 und LESEN 8-9 berücksichtigen die aufgrund theoretischer Überlegungen und empirischer Vorbefunde als am bedeutsamsten anzusehenden Aspekte des Leseverständnisses (vgl. Kap. 4). Dass die Tests jeweils aus einem Subtest zur Erfassung der basalen Lesekompetenz und einem Subtest zur Erfassung des Textverständnisses bestehen, ent-

spricht dem aktuellen Stand der Forschung, demgemäß sich Leseverständnis aus diesen Komponenten zusammensetzt (vgl. z. B. Schneider, 2008).

Subtest BLK. Mit der Satzleseaufgabe wird die basale Lesekompetenz, also die Lesegeschwindigkeit und -genauigkeit sowie das Verständnis einfacher, kurzer Sätze überprüft. Die Erfassung der basalen Lesekompetenz über diese Art der Operationalisierung ist offensichtlich, da sie schnelles und korrektes Lesen erfordert. Somit weist der Subtest BLK Augenscheinvalidität auf. Die Überprüfung der technischen Lesefertigkeit auf diese Weise hat sich auch bereits bei anderen Tests bewährt (vgl. z. B. SLS 5-8 von Auer et al., 2005, Kap. 5.3.4). Der Fokus des Subtests BLK liegt dabei vor allem auf der Lesegeschwindigkeit, welche sich in der Sekundarstufe als zuverlässiger Prädiktor für das Leseverständnis erwiesen hat.

Auf Basis der Normdaten wurden zur weiteren Prüfung der Inhaltsvalidität für den Subtest BLK die Fehler und Auslassungen, die in Kapitel 15 auf Itemebene durchgeführt wurden, auf Personenebene betrachtet. Dadurch sollte zum einen geprüft werden, ob die Schüler die Items tatsächlich in gewünschter Weise bearbeiteten oder ob sie andere, unerwünschte Strategien anwendeten (z. B. willkürliches Ankreuzen ohne Lesen der Sätze). Hierfür wurde betrachtet, wie viele korrekte Lösungen, Fehler und nicht in Angriff genommene Items für die einzelnen Schüler vorliegen. Die Auswertung erfolgte über alle Schüler der Normstichprobe von der sechsten bis zur neunten Klassenstufe ($N = 2\,584$). Im Mittel kamen die Schüler auf 57.70 ($SD = 15.00$) korrekt gelöste Items, 1.63 ($SD = 3.34$) Fehler bzw. Auslassungen und 40.67 ($SD = 15.16$) nicht in Angriff genommene Items. Die Tabellen 49 bis 51 in Anhang E zeigen die Häufigkeiten der einzelnen Rohwerte (Rohwert = Summe korrekt bearbeiteter Items), die vorkommenden Fehlerhäufigkeiten und die Häufigkeiten für die Summe nicht bearbeiteter Items. Dass die Schüler im Mittel nur ein bis zwei Items falsch lösten bzw. nur ein bis zwei Sätze ausließen, spricht für eine sorgfältige Arbeitsweise und gründliches Lesen beim größten Teil der Schüler.

Zusammen mit den Ergebnissen zum Subtest BLK in Kapitel 15 ergibt sich also, dass sowohl auf Itemebene als auch auf Personenebene wenige Fehler und Auslassungen vorkamen. Dies spricht zum einen für eine angemessene Leichtigkeit der Sätze und dafür, dass sie tatsächlich basale Lesekompetenzen erfassen, ohne zusätzlich hierarchiehöhere Verständnisprozesse zu erfordern. Zum anderen spricht dies dafür, dass die Schüler die Items in gewünschtem Sinne sorgfältig bearbeiteten. Insgesamt deuten die Ergebnisse sowie die theoretischen Einschätzungen auf eine hohe inhaltliche Validität des Subtests BLK hin.

Subtest TV. Für die inhaltliche Validität des Subtests TV spricht die stringent aus der Theorie abgeleitete Aufgabenstellung und Itemgenerierung. Mithilfe von Texten verschiedener Genres und den dazu gestellten Verständnisfragen unterschiedlichen Schwierigkeitsgrades wird das Textverständnis, einschließlich der Fähigkeit zur Bil-

dung von Kohärenzen und Inferenzen sowie der Textverarbeitung auf den Metaebenen, umfassend abgebildet. Die Subtests TV folgen dabei theoretisch motivierten Modellvorstellungen des Leseprozesses (s. z. B. Richter et al., 2002). Der Inhalt wurde auch von Experten als inhaltlich valide zur Erfassung der hierarchiehöheren kognitiven Aspekte der Lesekompetenz beurteilt (vgl. 13.3.2 bis 13.3.4) und bei der Itemselektion wurde stets die inhaltliche Bedeutsamkeit der Items für die angemessene Abbildung des Konstruktes berücksichtigt.

17.3 Konstruktvalidität

Die theoretische Annahme der Eindimensionalität des Subtests TV wird von den Ergebnissen der Itemanalysen auf Basis der Normdaten (s. Kap. 15) für beide Tests empirisch gestützt. Dies kann bereits als Hinweis auf Konstruktvalidität gewertet werden, da dies mit dem aktuellen Forschungsstand zum Leseverständnis konform geht (vgl. Kap. 4.3). Darüber hinaus zeigt sich, dass die beiden Subtests nur in mittlerer Höhe (LESEN 6-7: r = .45; LESEN 8-9: r = .40) korrelieren, was ebenfalls dem aktuellen Kenntnisstand der Forschung entspricht, dem zufolge die basale Lesekompetenz und das Leseverständnis zwar nicht unabhängig voneinander sind, aber auch nicht hoch korrelieren (vgl. Schneider, 2008).

Zur weiteren Überprüfung der Konstruktvalidität beider Subtests und des Gesamtergebnisses wurden zum einen die Ergebnisse von LESEN 6-7 und LESEN 8-9 mit den Ergebnissen des LGVT 6-12 und des WLST 7-12 korreliert. Da diese Tests Lesegeschwindigkeit und Leseverständnis bzw. das damit in Zusammenhang stehende Wissen über Lesestrategien erfassen, wurde jeweils eine substanzielle Korrelation erwartet. Zum anderen wurde geprüft, ob LESEN 6-7 und LESEN 8-9 im Hinblick auf aus der Theorie und auf Basis von empirischen Vorbefunden abgeleitete Hypothesen erwartungskonforme Ergebnisse generieren.

17.3.1 Hypothesen

Im Folgenden werden die im Rahmen der Prüfung der Konstruktvalidität getesteten Hypothesen dargestellt. Zuvor werden jeweils die den Hypothesen zugrunde liegenden theoretischen Überlegungen bzw. empirischen Befunde aus dem ersten Teil der Arbeit kurz zusammengefasst.

Korrelation mit anderen Lesetestergebnissen. Im Sinne konvergenter Validität sollten die Ergebnisse von LESEN 6-7 und LESEN 8-9 mit den Ergebnissen anderer Lesetests, die dasselbe Konstrukt erfassen, substanziell korrelieren. Der LGVT 6-12 enthält ähnlich wie LESEN 6-7 und LESEN 8-9 eine Skala für die Lesegeschwindigkeit (LGVT-LG) und eine Skala für das Leseverständnis (LGVT-LV). Aus diesem Grund wurde eine

substanzielle Korrelation der Subtestergebnisse von LESEN 6-7 und LESEN 8-9 mit den Ergebnissen beider Skalen des LGVT 6-12 erwartet.

Weiter wurde angenommen, dass die Ergebnisse der Subtests BLK von LESEN 6-7 und LESEN 8-9 höher mit den Ergebnissen der Skala LGVT-LG korrelieren als mit den Ergebnissen der Skala LGVT-LV, da der Subtest BLK und die Skala LGVT-LG beide basale Lesekompetenzen erfassen sollen. Für den Subtest TV wurde dagegen das Umgekehrte erwartet, da der Subtest TV und die Skala LGVT-LV beide das Leseverständnis erfassen sollen. Diesen Überlegungen entsprechend wurden folgende Hypothesen aufgestellt:

H1a: Die Ergebnisse beider Subtests sowie die Gesamtergebnisse von LESEN 6-7 und LESEN 8-9 korrelieren substanziell mit den Ergebnissen beider Skalen des LGVT 6-12.

H1b: Die Ergebnisse des Subtests BLK von LESEN 6-7 und LESEN 8-9 korrelieren jeweils höher mit den Ergebnissen der Skala LGVT-LG als mit den Ergebnissen der Skala LGVT-LV, während die Ergebnisse des Subtests TV von LESEN 6-7 und LESEN 8-9 jeweils höher mit den Ergebnissen der Skala LGVT-LV korrelieren als mit den Ergebnissen der Skala LGVT-LG.

Korrelation mit den Ergebnissen von Lesestrategie-Wissenstests. Wie bereits in Kapitel 7 deutlich wurde, spielt Strategiewissen beim Leseverständnis eine Rolle (D. H. Rost & Buch, 2010; Wellman, 1983). Selbst bei Kontrolle von Dekodierfähigkeit, kognitiver Grundfähigkeit, verbalem Selbstkonzept und Leseinteresse verbessert Strategiewissen die Vorhersage der Leseleistung signifikant. Das Lesestrategiewissen war bei PISA 2000 der zweitbeste Prädiktor der Leseleistung (Artelt et al., 2002). Bei der Validierung des WLST 7-12 zeigte sich ein deutlicher Zusammenhang mit den Ergebnissen des Leseverständnistests von PISA 2003 ($r = .48$, $p < .01$, $N = 971$; Schlagmüller & Schneider, 2007). Daher wird angenommen, dass die Ergebnisse der Lesestrategie-Wissenstests (des MKT in der sechsten Klassenstufe und des WLST 7-12 in den Klassenstufen sieben bis neun) substanziell mit den Ergebnissen von LESEN 6-7 und LESEN 8-9 korrelieren.

Die Anwendung von Lesestrategien hat sich insbesondere für hierarchiehöhere Leseprozesse als bedeutsam erwiesen (vgl. Artelt et al., 2007; W. Lenhard, 2013). Bei der Validierung des LGVT 6-12 korrelierten die Ergebnisse der Skala LGVT-LV deutlich höher mit den Ergebnissen im WLST 7-12 ($r = .46$) als die Ergebnisse der Skala LGVT-LG ($r = .27$), wenngleich beide Korrelationen signifikant ausfielen ($p < .01$, $N = 809$; Schneider et al., 2007). Daher sollte die Korrelation der Ergebnisse in den Lesestrategie-Wissenstests auch jeweils mit den Ergebnissen des Subtests TV von LESEN 6-7 und LESEN 8-9 höher korrelieren als mit den Ergebnissen des Subtests BLK. Die entsprechenden Hypothesen lauten somit:

H2a: Die Ergebnisse der Lesestrategie-Wissenstests korrelieren substanziell mit den Ergebnissen beider Subtests und den Gesamtergebnissen von LESEN 6-7 und LESEN 8-9.

H2b: Die Korrelation der Ergebnisse in den Lesestrategie-Wissenstests mit den Ergebnissen im Subtest TV von LESEN 6-7 und LESEN 8-9 fällt höher aus als die Korrelation mit den Ergebnissen im Subtest BLK.

Klassenstufenunterschiede. Verschiedene Studien zeigten, dass Schüler auch in der Sekundarstufe noch an Lesekompetenz hinzugewinnen – wenngleich der Zugewinn nicht mehr so groß ausfällt wie in der Grundschule und er im Verlauf der Sekundarschuljahre weiter abflacht (vgl. W. Lenhard, 2013; Philipp, 2011b). Das Ausmaß an Leistungszuwachs scheint für die ersten Sekundarschuljahre bei etwa einer drittel bis einer halben Standardabweichung pro Schuljahr zu liegen und im weiteren Verlauf bis auf etwa 0.2 oder 0.1 Standardabweichungen abzunehmen (z. B. Retelsdorf & Möller, 2008; Baumert & Artelt, 2002; R. Lehmann & Lenkeit, 2008; Philipp, 2011b). Die Automatisierung basaler Leseprozesse wird in den Sekundarschuljahren noch gesteigert, und der Aufbau eines Strategierepertoires sowie die gezielte Anwendung von Lesetechniken und Strategien zum verstehenden Lesen finden häufig erst im Verlauf der Sekundarschuljahre statt und befördern das sinnentnehmende Lesen (vgl. Artelt et al., 2007; Schneider, 2008). Daher wird auch für die Normstichproben von LESEN 6-7 und LESEN 8-9 erwartet, dass Schüler der jeweils höheren Klassenstufe in beiden Subtests bessere Leseleistungen zeigen als Schüler der jeweils niedrigeren Klassenstufe. Da die hier vorliegenden Daten aus einer Querschnittserhebung stammen, können sich die Hypothesen nur auf Leistungsunterschiede zwischen den Klassenstufen beziehen; die Prüfung von Leistungszuwächsen im eigentlichen Sinne ist nicht möglich. Die Hypothese bezüglich der Klassenstufenunterschiede lautet daher wie folgt:

H3: Sowohl bei LESEN 6-7 als auch bei LESEN 8-9 erreichen Schüler der jeweils höheren Klassenstufe (also 7 bzw. 9) in beiden Subtests höhere Ergebniswerte als Schüler der jeweils niedrigeren Klassenstufe (also 6 bzw. 8).

Wie erläutert, fand für das Gesamtergebnis kein Klassenstufenvergleich statt.

Unterschiede zwischen den Schularten. Bezüglich der verschiedenen Regelschularten in Deutschland zeigten sich immer wieder große Unterschiede in der Leseleistung (vgl. Kap. 6). Zwar waren bei PISA 2006 in allen Schularten alle Leseniveaustufen vertreten, jedoch jeweils mit sehr unterschiedlichen Anteilen. So erreichten beispielsweise knapp 4 % der Hauptschüler, etwa 25 % der Realschüler und etwa 70 % der Gymnasiasten die höchsten Niveaustufen IV und V. Bei PISA 2009 zeichnete sich ein ähnliches Bild ab, wobei die hier am oberen Ende hinzugefügte Niveaustufe VI von Hauptschülern nicht erreicht wurde und am unteren Ende keine Gymnasiasten unterhalb der neu gebildeten Niveaustufe Ib lagen (Naumann et al., 2010). Weiter lagen bei PISA 2009 die Lesetestergebnisse der deutschen Hauptschüler etwa zwei Standardabweichungen unter den Ergebnissen der Gymnasiasten. Die Ergebnisse der Realschüler und der Schüler integrierter Gesamtschulen lagen dazwischen. Insgesamt unterschieden sich die Schularten also trotz Überlappungsbereichen deutlich. Ein entsprechen-

des Ergebnismuster wurde daher auch für LESEN 6-7 und LESEN 8-9 jeweils für beide Subtests und das Gesamtergebnis erwartet. Es wurde somit folgende Hypothese aufgestellt:

H4: Gymnasiasten erreichen sowohl bei LESEN 6-7 als auch bei LESEN 8-9 in beiden Subtests und bezüglich des Gesamtergebnisses die höchsten Ergebniswerte, Realschüler liegen im Mittelfeld und Hauptschüler erreichen die niedrigsten Ergebniswerte.

Geschlechterunterschiede. In zahlreichen – wenn auch nicht in allen – Studien zeigen sich Geschlechterunterschiede hinsichtlich der Leseleistung (vgl. Kap. 6). Dort, wo sie auftreten, fallen sie stets zugunsten der Mädchen aus. Beispielsweise wiesen in den PISA-Studien Mädchen kontinuierlich in allen Bereichen der Lesekompetenz bessere Leistungen auf als Jungen (Naumann et al., 2010). Die Unterschiede fielen stets signifikant aus, wenngleich die Effektstärken klein waren. Bei DESI zeigte sich ebenfalls bei der Lesekompetenz ein Vorteil der Mädchen gegenüber den Jungen, und auch bei den Normierungen der Tests LGVT 6-12 und SLS 5-8 erzielten die Mädchen tendenziell bessere Ergebnisse als die Jungen (Schneider et al., 2007; Auer et al., 2005; DESI-Konsortium, 2006). Daher wurde auch für LESEN 6-7 und LESEN 8-9 erwartet, dass die Mädchen den Jungen überlegen sind – allerdings wurde entsprechend der Vorbefunde mit eher kleinen Effekten gerechnet. Es ergab sich daher folgende Hypothese:

H5: Die Mädchen erzielen sowohl bei LESEN 6-7 als auch bei LESEN 8-9 in beiden Subtests und im Gesamtergebnis höhere Ergebniswerte als die Jungen.

Unterschied zwischen Schülern mit und ohne LRS-Diagnose. LRS wird unter anderem über schwache Leseleistungen definiert (vgl. Kap. 5.1). Daher sollten in einem Test, der das Leseverständnis erfasst, Schüler mit einer LRS-Diagnose im Mittel schwächere Leistungen zeigen als Schüler ohne LRS-Diagnose (vgl. auch Petermann & Daseking, 2012). Allerdings wurde auch hier nicht mit allzu großen Effekten gerechnet, da LRS im Sekundarschulalter hauptsächlich in schwachen Rechtschreibleistungen zum Ausdruck kommt, wohingegen sich die Leseleistung bis zum Erreichen der mittleren und höheren Klassenstufen weitgehend normalisiert hat. Für LESEN 6-7 und LESEN 8-9 wurde entsprechend folgende Hypothese aufgestellt:

H6: Schüler mit LRS-Diagnose erzielen sowohl bei LESEN 6-7 als auch bei LESEN 8-9 in beiden Subtests sowie in Bezug auf das Gesamtergebnis niedrigere Ergebniswerte als Schüler ohne LRS-Diagnose.

Unterschied zwischen Schülern mit Deutsch als Muttersprache und Schülern mit einer anderen Muttersprache als Deutsch. Die Sprachgebundenheit des Leseverständnisses wurde eingangs bereits angesprochen (s. Kap. 1), und auf den Einfluss

eines Migrationshintergrunds sowie auf Unterschiede zwischen Schülern mit Deutsch als Muttersprache und Schülern mit einer anderen Muttersprache als Deutsch wurde bereits ausführlicher eingegangen (s. Kap. 6 und Kap. 7). Schüler mit Migrationshintergrund bzw. Schüler, die mit einer anderen Muttersprache als Deutsch aufwachsen, schnitten in zahlreichen Studien zur Lesekompetenz deutlich schlechter ab als Schüler mit Deutsch als Muttersprache bzw. Schüler ohne Migrationshintergrund, und sie gelten häufig als Risikogruppe für Leseprobleme (vgl. DESI-Konsortium, 2006; Artelt et al., 2007; Bos et al., 2003; Baumert & Schümer, 2001). Allerdings zeigte sich bei genauerer Betrachtung, dass die Unterschiede lediglich in Bezug auf das Textverständnis bestanden, nicht jedoch in Bezug auf die Leseflüssigkeit auf Wortniveau, und in Bezug auf das Lesen mehrerer Sätze zeigten sich ebenfalls nur teilweise Unterschiede (Limbird & Stanat, 2006; A. E. Marx & Stanat, 2011). Da jedoch in der vorliegenden Arbeit für beide Subtests ein Mindestmaß an Leseverständnis erforderlich ist, wurde aufgrund der Sprachgebundenheit von Leseverständnis und im Sinne der Vorbefunde angenommen, dass Schüler mit Deutsch als Muttersprache bei LESEN 6-7 und LESEN 8-9 bessere Testergebnisse erzielen als Schüler mit einer anderen Muttersprache als Deutsch, was folgendermaßen als Hypothese formuliert wurde:

H7: Schüler mit einer anderen Muttersprache als Deutsch erzielen in beiden Subtests sowie in Bezug auf das Gesamtergebnis bei LESEN 6-7 und LESEN 8-9 niedrigere Ergebniswerte als Schüler mit Deutsch als Muttersprache.

17.3.2 Methode

Zur Prüfung der Hypothesen H1a/b und H2a/b, die sich auf Korrelationen der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit Ergebnissen anderer Tests beziehen, wurden Korrelationsanalysen durchgeführt. Zur Entscheidung über die Hypothesen H3 bis H7, die Mittelwertsunterschiede zwischen verschiedenen Subgruppen der Norm- bzw. Validierungsstichproben postulierten, wurden *t*-Tests und Varianzanalysen eingesetzt bzw. deren entsprechende nonparametrische Pendants.

Korrelationen (H1a/b und H2a/b). Die Korrelationen der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit Ergebnissen anderer Tests erfolgten auf Basis der Daten der Validierungsstichprobe, da die anderen Tests im Rahmen der Normierung nicht durchgeführt wurden. Es wird angenommen, dass die Ergebnisse aller eingesetzter Tests intervallskaliert sind, daher wurde bei Normalverteilung der Ergebniswerte beider Tests die Produkt-Moment-Korrelation nach Pearson angewendet, bei einer Abweichung der Daten von der Normalverteilung bei einem oder bei beiden Tests wurde die Rangkorrelation nach Spearman herangezogen. Die Prüfung auf Normalverteilung erfolgte mittels KS-Tests. Bei nicht signifikantem Ergebnis konnte von normalverteilten Daten ausgegangen werden (Ergebnisse der KS-Tests s. Tab. 52 in Anhang E). Die Prüfung erfolgte mit einem Signifikanzniveau von $\alpha = .20$, da die Nullhypothese, dass die Daten

normalverteilt sind, beibehalten und somit der β-Fehler (also eine fälschliche Annahme der Normalverteilung) klein gehalten werden sollte. Die Bewertung der Korrelationshöhe zwischen den Ergebnissen der Tests folgt den in Kapitel 5.3.2.1 genannten Richtlinien von Fisseni (2004).

Da die Höhe der Korrelationen der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit Ergebnissen anderer Tests auch von der Reliabilität der anderen Tests beeinflusst wird, wurde eine den Messfehler der anderen Tests berücksichtigende, einfache Minderungskorrektur gemäß der Beschreibung von Lienert und Raatz (1998, S. 257) durchgeführt. Anhand der einfachen Minderungskorrektur wird geschätzt, wie hoch die Ergebnisse von LESEN 6-7 bzw. LESEN 8-9 mit den Ergebnissen der anderen Tests korrelieren würden, wenn diese anderen Tests perfekt reliabel wären. Hierfür wurden die Retest-Reliabilitätswerte herangezogen. Da für den in den sechsten Klassen durchgeführten MKT keine Retestreliabilitätswerte vorlagen, erfolgte hier keine Minderungskorrektur.

Bei der Validierungsstichprobe handelte es sich um eine im Vergleich zur Normstichprobe kleinere und weniger repräsentative Stichprobe. Dies ging mit einer geringeren Varianz der Rohwerte bei der Validierung im Vergleich zur Normstichprobe einher (Varianzverhältnis Normierung : Validierung bei LESEN 6-7 für BLK 200.51 : 189.06, für TV 47.75 : 42.12; bei LESEN 8-9 für BLK 243.83 : 158.26, für TV 50.84 : 40.70). Da sich eine geringere Streuung der Werte negativ auf die Höhe der Korrelationen auswirkt, wurde die von Lienert und Raatz (1998, S. 266) für diesen Fall empfohlene Korrekturformel zur Bestimmung der Validitätswerte angewendet. Diese berücksichtigt die vorliegende Varianzeinschränkung bei den Daten der Validierungsstichprobe im Vergleich zur Normstichprobe und verhindert eine mit der Varianzeinschränkung einhergehende Unterschätzung der tatsächlichen Korrelationshöhe[10].

Hypothesentests auf Basis von Korrelationen setzen eigentlich normalverteilte Daten voraus (Bortz & Schuster, 2010, S. 161f.). Da zum Teil die Normalverteilungsannahme verletzt ist (s. Tab. 52 in Anhang E), wurde zur Prüfung auf signifikant von Null abweichende Korrelationen ein von Bortz und Schuster (2010, S. 162) beschriebener Signifikanztest eingesetzt, der robust gegenüber einer Verletzung der Verteilungsannahme ist.

Die Prüfung auf signifikante Unterschiede zwischen Korrelationen erfolgte wie von Bortz und Schuster (2010, S. 166) beschrieben mithilfe des Fisher-*Z*-Tests für abhängige Stichproben. Wenn der Betrag des Ergebniswertes größer als ein kritischer *z*-Wert ist, ist der Unterschied zwischen den Korrelationen signifikant. Da gerichtete Hypothesen getestet wurden, erfolgte die Testung einseitig, und somit lag der kritische Wert bei einem Alpha-Niveau von 5 % bei $z = 1.64$. Zur Beurteilung der praktischen Bedeutsamkeit signifikanter Effekte wird zusätzlich die Prüfgröße q nach Cohen (1988,

[10] Abweichungen der Validitätswerte von den in den Testmanualen angegebenen Werten sind auf diese Minderungskorrektur zurückzuführen, da diese bei den Berechnungen für die Testmanuale nicht vorgenommen wurde.

S. 110) verwendet, die die Differenz der Fisher-*Z*-transformierten Korrelationen darstellt und folgendermaßen bewertet werden kann: Ein Wert von $q = .10$ spricht für einen kleinen Effekt, ein Wert von $q = .30$ für einen moderaten Effekt und ein Wert von $q = .50$ für einen großen Effekt (Cohen, 1988, S. 113ff.).

Mittelwertsunterschiede (H3 bis H7). Im Folgenden wird nach einigen generellen Vorbemerkungen zur Auswertung im Hinblick auf Mittelwertsunterschiede zunächst das Vorgehen zur gemeinsamen Prüfung der Hypothesen H3 (Klassenstufenunterschiede), H4 (Schulartunterschiede) und H5 (Geschlechterunterschiede) beschrieben. Daraufhin wird das Vorgehen zur Prüfung von H6 (Unterschied zwischen Schülern mit und ohne LRS) dargestellt und schließlich das Vorgehen zur Prüfung von H7 (Unterschiede zwischen Schülern mit Deutsch als Muttersprache und Schülern mit einer anderen Muttersprache als Deutsch). Die Prüfung auf signifikante Mittelwertsunterschiede erfolgte abhängig von der Anzahl der unabhängigen Variablen (UVn) und abhängigen Variablen (AVn) sowie von der Erfüllung der Voraussetzungen jeweils varianzanalytisch, mithilfe von *t*-Tests oder mithilfe eines nonparametrischen Tests.

Die unter „Andere“ zusammengefassten Schularten sowie Schüler mit Deutsch und einer weiteren Muttersprache wurden wegen der jeweils sehr geringen Fallzahlen ausgeschlossen. Somit weist die UV Schulart drei Stufen (Hauptschule, Realschule und Gymnasium) und die UV Muttersprache zwei Stufen (Deutsch und Andere) auf. Bezüglich der UV Klassenstufe gibt es für jeden Test zwei Stufen (für LESEN 6-7 die Klassenstufen sechs und sieben, für LESEN 8-9 die Klassenstufen acht und neun). Auch die UV Geschlecht und die UV LRS sind jeweils zweistufig (Mädchen vs. Jungen bzw. LRS vs. keine LRS).

Zur Prüfung auf Klassenstufen-, Schulart- und Geschlechterunterschiede sowie auf Unterschiede zwischen Schülern mit und ohne LRS konnte die große Normstichprobe herangezogen werden. Für die Prüfung auf Unterschiede zwischen Schülern mit Deutsch als Muttersprache und Schülern mit einer anderen Muttersprache als Deutsch konnte nur die kleinere Validierungsstichprobe herangezogen werden. Die Klassenstufen-, Schulart- und Geschlechterunterschiede wurden für jeden Test in einer gemeinsamen Varianzanalyse geprüft. Für den Vergleich von Schülern mit LRS und Schülern ohne LRS wurden separate Analysen durchgeführt, wobei auch die Klassenstufenzugehörigkeit berücksichtigt wurde.

Für die Entscheidung, ob die Varianzanalysen multivariat oder univariat berechnet werden sollten, war die Korrelationshöhe der AVn zu betrachten. Laut Tabachnick und Fidell (2007, S. 268) eignet sich die multivariate Varianzanalyse (MANOVA) am besten für moderat korrelierte AVn. Dies trifft für beide Tests für die Subtests BLK und TV zu, die bei LESEN 6-7 zu $r = .45$ und bei LESEN 8-9 zu $r = .40$ korrelieren. Die Prüfung erfolgte daher multivariat mit den Ergebnissen der Subtests BLK und TV als AVn. Da sich das Gesamtergebnis aus den Subtestergebnissen zusammensetzt, wurden hierfür jeweils separate Analysen durchgeführt.

Generell wurde bei den Varianzanalysen aufgrund der durchweg ungleichen Zellbesetzungen die von Tabachnick und Fidell (2007, S. 217f.) empfohlene „Methode 1“ (entspricht bei SPSS der Standardeinstellung „Quadratsumme Typ III“) zurückgegriffen, bei der alle Zellen ungeachtet der tatsächlichen Stichprobengröße gleich gewichtet werden. Dabei handelt es sich um eine konservative Methode. Fälle mit fehlenden Werten für eine AV wurden wie bei Tabachnick und Fidell (2007, S. 277f.) beschrieben eliminiert. Zusätzlich zu den bereits ausgeschlossenen Ausreißern wurden daher für LESEN 6-7 drei Schüler ausgeschlossen, weil sie keinen BLK-Wert hatten. Weitere 28 Schüler wurden ausgeschlossen, weil ihr TV-Wert fehlte. Für LESEN 8-9 wurden zwei Schüler ausgeschlossen, weil ihr BLK-Wert fehlte. So verblieben für die Analysen 1 551 Schüler bei LESEN 6-7 und 929 Schüler bei LESEN 8-9.

Gerichtete Hypothesen wurden einseitig getestet, indem die ermittelten *p*-Werte halbiert wurden. Da statistische Tests lediglich eine Wahrscheinlichkeitsaussage über das Zutreffen der Null- oder Alternativhypothese treffen, wurde zur Bestimmung der praktischen Bedeutsamkeit von Unterschieden für jeden signifikanten Unterschied die bereits erläuterte Effektstärke d nach Cohen (1988, S. 20ff.) berechnet (vgl. Kap. 16). Darüber hinaus wurde das partielle Eta-Quadrat (η_p^2) bestimmt. Dieses gibt an, welcher Varianzanteil durch den Effekt aufgeklärt werden kann, wenn alle anderen Effekte auspartialisiert sind. Nach Cohen (1988, S. 284ff.) wird die Höhe dieser Werte wie folgt interpretiert: Werte ab $\eta_p^2 = .01$ sprechen für einen kleinen Effekt, Werte ab $\eta_p^2 = .06$ für einen mittelgroßen Effekt und Werte ab $\eta_p^2 = .14$ für einen großen Effekt.

Zur Prüfung der Hypothesen bezüglich *Klassenstufen-, Schulart- und Geschlechterunterschieden* (H3 bis H5) wurde für die Subtests von LESEN 6-7 und LESEN 8-9 eine MANOVA mit den zwei AVn BLK und TV und den UVn Klassenstufe, Schulart und Geschlecht durchgeführt. Bezüglich des Gesamtergebnisses wurden, wie bereits erwähnt, keine Klassenstufenvergleiche durchgeführt. Für das Gesamtergebnis wurde daher pro Klassenstufe eine ANOVA durchgeführt mit dem Gesamtergebnis von LESEN 6-7 bzw. LESEN 8-9 als AV sowie Schulart und Geschlecht als UVn.

Die Berechnung von ANOVAs setzt Unabhängigkeit der Stichproben, Normalverteilung der Daten in allen Stichproben sowie Gleichheit der Varianzen in allen Stichproben (Varianzhomogenität) voraus (Bortz & Schuster, 2010, S. 212ff.). Unabhängigkeit der Stichproben kann aus dem Untersuchungsdesign geschlossen werden und liegt vor, wenn die Faktorstufen in verschiedenen Stichproben erhoben wurden. Die Normalverteilungsvoraussetzung kann anhand von KS-Tests überprüft werden, die Varianzhomogenität mithilfe von Levene-Tests. Da bei beiden Tests die Beibehaltung der Nullhypothese erwünscht ist, ist jeweils auf einem Niveau von $\alpha = .20$ zu testen. Da die ANOVA bei ungleicher Zellbesetzung nicht unbedingt robust auf eine Verletzung ihrer Voraussetzungen reagiert, sind die Voraussetzungen zu prüfen. Vor allem bei kleinen Stichproben ist eine Voraussetzungsverletzung bei ungleich großen Stichproben kritisch (Bortz & Schuster, 2010, S. 214).

Die Berechnung von MANOVAs setzt über die Annahmen der ANOVA hinaus eine multivariate Normalverteilung der AVn in jeder Gruppe und homogene Kovarianz-Matrizen für die einzelnen AVn voraus (Bortz & Schuster, 2010, S. 481). Die Voraussetzung der multivariaten Normalverteilung lässt sich anhand eines von DeCarlo (1997) beschriebenen Schiefe- und Exzess-Tests prüfen. Dies ist jedoch nur notwendig, wenn eine univariate Normalverteilung gegeben ist, da diese eine notwendige – aber nicht hinreichende – Bedingung für das Vorliegen einer multivariaten Normalverteilung darstellt (vgl. DeCarlo, 1997). Mit dem Box-Test kann die Homogenität der Kovarianz-Matrizen geprüft werden. Dabei ist jedoch zu berücksichtigen, dass der Box-Test – insbesondere bei einer Verletzung der Annahme multivariater Normalverteilung – sehr sensitiv ist (vgl. Bortz & Schuster, 2010, S. 500). Da auch hier die Beibehaltung der Nullhypothese erwünscht ist, sollte ebenfalls auf einem Niveau von $\alpha = .20$ getestet werden. Bei großen Stichproben ist eine Verletzung der Voraussetzungen der MANOVA jedoch praktisch zu vernachlässigen, wenn die Stichproben gleich groß sind (Ito, 1969; Ito & Schul, 1964; Stevens, 2002). Laut Mardia (1971) sollte jedoch eine Stichprobengröße von $n = 20$ pro Zelle bereits ausreichen, um die Robustheit des Verfahrens auch bei ungleichen Zellbesetzungen sicherzustellen. Insgesamt ist davon auszugehen, dass die MANOVA bei größeren Stichproben ein robustes und teststarkes Verfahren ist (vgl. Stevens, 2002, S. 277).

Die kleinste Zellbesetzung für die MANOVA zur Prüfung auf Unterschiede zwischen den Klassenstufen (H3), zwischen den Schularten (H4) und zwischen den Geschlechtern (H5) in Bezug auf die Subtests lag für LESEN 6-7 bei $n = 63$ (s. Tab. 53 in Anhang E). Somit waren alle Zellen mit $n > 20$ besetzt, was für die Robustheit der Analyse gegenüber Verletzungen der Normalverteilungsvoraussetzung selbst bei ungleichen Zellbesetzungen spricht, und deshalb mussten keine KS-Tests durchgeführt werden. Der Box-Test auf Gleichheit der Kovarianz-Matrizen fiel wie erwartet signifikant aus ($F(33, 2541138) = 3.06$, $p < .01$). Der Levene-Test auf Gleichheit der Fehlervarianzen wurde ebenfalls für beide Subtests signifikant (BLK: $F(11, 1\,497) = 3.61$, $p < .01$; TV: $F(11, 1\,497) = 1.77$, $p = .05$). Somit war keine Varianzhomogenität gegeben.

Für LESEN 8-9 lag die kleinste Zellbesetzung für die MANOVA bei $n = 35$ (s. Tab. 54 in Anhang E). Somit waren auch hier alle Zellen mit $n > 20$ besetzt, was wieder für die Robustheit der Analyse gegenüber Verletzungen der Normalverteilungsvoraussetzung selbst bei ungleichen Zellbesetzungen spricht, weshalb auch hier keine KS-Tests durchgeführt werden mussten. Der Box-Test auf Gleichheit der Kovarianz-Matrizen wurde erwartungsgemäß signifikant ($F(33, 7071171.7) = 2.19$, $p < .01$). Für LESEN 8-9 wurde auch der Levene-Test wieder für beide Subtests signifikant (BLK: $F(11, 885) = 1.68$, $p = .07$; TV: $F(11, 885) = 3.33, p < .01$). Somit war Varianzhomogenität für keinen Subtest gegeben.

Insgesamt sind somit die Voraussetzungen für eine MANOVA weder für LESEN 6-7 noch für für LESEN 8-9 erfüllt. Da jedoch ausreichend große Zellbesetzungen vorliegen und somit von einer Robustheit des Verfahrens ausgegangen werden kann, wird

dennoch das parametrische Vorgehen gegenüber dem nonparametrischen Vorgehen bevorzugt. Aufgrund der Voraussetzungsverletzungen wird allerdings auf die robuste Prüfgröße Pillai-Spur zurückgegriffen.

In Bezug auf das Gesamtergebnis wurde pro Klassenstufe eine ANOVA berechnet. Für die sechste Klassenstufe lag die kleinste Zellbesetzung bei $n = 63$, für die siebte Klassenstufe bei $n = 119$, für die achte Klassenstufe bei $n = 55$ und für die neunte Klassenstufe bei $n = 35$ (s. Tab. 53 und 54 in Anhang E). Die KS-Tests fallen für keine Klassenstufe signifikant aus (s. Tab. 55 in Anhang E), was bedeutet, dass die Normalverteilung für alle Klassenstufen gegeben ist. Der Levene-Test erwies sich für die sechste und siebte Klassenstufe als nicht signifikant (6. Kl.: $F < 1$; 7. Kl.: $F(5, 831) = 1.07$, n.s.), für die achte und neunte Klasse dagegen als signifikant (8. Kl.: $F(5, 503) = 2.68$, $p = .02$; 9. Kl. $F(5, 382) = 2.00$, $p = .08$). Somit konnte die Voraussetzung der Gleichheit der Fehlervarianzen für beide Klassenstufen von LESEN 6-7 angenommen werden, für beide Klassenstufen von LESEN 8-9 nicht. Da die Stichproben jedoch nicht sehr klein sind, wurde dennoch auch für die achte und die neunte Klassenstufe jeweils eine ANOVA durchgeführt.

Da die UV Schulart mehr als zwei Stufen umfasste, wurden bei Feststellung signifikanter Mittelwertsunterschiede zusätzlich Post-Hoc-Tests in Form multipler paarweiser Vergleiche durchgeführt, um herauszufinden, zwischen welchen Schularten ein signifikanter Unterschied besteht. Hier wurde bei gegebener Varianzhomogenität der Scheffé-Test berechnet, der das Signifikanzniveau so festlegt, dass alle möglichen Linearkombinationen von Gruppenmittelwerten getestet werden können, was ihn sehr konservativ macht. Bei Verletzung der Varianzhomogenität wurde der Tamhane-T2-Test durchgeführt (ebenfalls ein konservativer paarweiser Vergleichstest), welcher auf einem *t*-Test basiert.

Für den *Vergleich von Schülern mit vs. ohne LRS* (H6) wurde für die Subtestergebnisse für beide Tests eine MANOVA durchgeführt mit den AVn BLK und TV sowie den UVn LRS vs. kein LRS und Klassenstufe (6 vs. 7 bzw. 8 vs. 9). Für die Unterschiedsprüfung im Hinblick auf das Gesamtergebnis von LESEN 6-7 und LESEN 8-9 wurde auf Klassenstufenebene jeweils ein *t*-Test mit dem Gesamtergebnis als AV und LRS vs. kein LRS als UV durchgeführt.

Bei LESEN 6-7 war die kleinste Zellbesetzung für die MANOVA mit $n = 79$ relativ groß, und somit waren alle Zellen mit $n > 20$ ausreichend stark besetzt, um eine Robustheit der MANOVA gegenüber Verletzungen der Normalverteilungsvoraussetzung auch bei ungleichen Zellbesetzungen zu gewährleisten (s. Tab. 56). Ein Durchführen von KS-Tests war daher nicht notwendig. Der Box-Test auf Gleichheit der Kovarianz-Matrizen erwies sich bei einem Alpha-Niveau von .20 als signifikant ($F(9, 396195.4) = 1.48$, $p = .15$). Dies war jedoch bei großen Stichproben zu erwarten, da der Box-Test sehr sensitiv ist. Der Levene-Test auf Gleichheit der Fehlervarianzen stellte sich für den Subtest BLK als nicht signifikant heraus ($F < 1$), für den Subtest TV dagegen als signifikant ($F(3, 1581) = 2.81$, $p = .04$). Somit war Varianzhomogenität nur für den

Subtest BLK gegeben. Da nicht alle Voraussetzungen erfüllt waren, wurde die robuste Prüfgröße Pillai-Spur verwendet.

Bei LESEN 8-9 betrug die kleinste Zellbesetzung $n = 25$ (s. Tab. 56). Somit waren auch hier alle Zellen mit $n > 20$ besetzt, was für eine Robustheit der MANOVA gegenüber Verletzungen der Normalverteilungsvoraussetzung spricht. Auch hier war daher ein Durchführen von KS-Tests nicht notwendig. Der Box-Test auf Gleichheit der Kovarianz-Matrizen fiel bei einem Alpha-Niveau von .20 auch für LESEN 8-9 signifikant aus ($F(9, 46671.67) = 1.63$, $p = .10$). Gleichheit der Kovarianz-Matrizen war somit nicht gegeben. Der Levene-Test auf Gleichheit der Fehlervarianzen fiel für den Subtest BLK signifikant aus ($F(3, 924) = 4.13, p = .01$), für den Subtest TV nicht ($F < 1$). Somit ist Varianzhomogenität nur für den Subtest TV gegeben, für den Subtest BLK nicht. Da nicht alle Voraussetzungen erfüllt waren, wurde auch hier die robuste Prüfgröße Pillai-Spur verwendet.

Für das Gesamtergebnis wurde pro Klassenstufe ein *t*-Test berechnet, um zu prüfen, ob sich die Gesamtergebnisse der Gruppe der Schüler mit einer LRS-Diagnose signifikant von den Gesamtergebnissen der Gruppe der Schüler ohne LRS-Diagnose unterscheiden. Voraussetzungen für einen *t*-Test für unabhängige Stichproben sind eine Unabhängigkeit der Messwerte verschiedener Personen, Intervalldatenniveau der Messwerte, eine Normalverteilung der AV-Werte in beiden Gruppen sowie Varianzhomogenität. Die Unabhängigkeit der Messwerte verschiedener Personen ergab sich aus dem Design, die Normalverteilungsvoraussetzung wurde anhand von KS-Tests überprüft und die Varianzhomogenität anhand von Levene-Tests. Der KS-Test fiel in keiner Klassenstufe für irgendeine Subgruppe signifikant aus (6. Kl. LRS: $KS - Z = 0.79$, keine LRS: $KS - Z = 0.62$; 7. Kl. LRS: $KS - Z = 0.79$, keine LRS: $KS - Z = 0.44$; 8. Kl. LRS: $KS - Z = 0.85$, keine LRS: $KS - Z = 0.83$; 9. Kl. LRS: $KS - Z = 0.94$, keine LRS: $KS - Z = 0.65$, alle $p > .20$). Auch der Levene-Test fiel in keiner Klassenstufe für irgendeine Subgruppe signifikant aus (6., 7. und 9. Kl.: $F < 1$; 8. Kl.: $F = 1.48$). Somit waren die Voraussetzungen für die Berechnung von *t*-Tests erfüllt.

Für die Prüfung der *Unterschiede zwischen Schülern mit Deutsch als Muttersprache und Schülern mit einer anderen Muttersprache als Deutsch* (H7) musste zunächst festgestellt werden, welche Variablen bei der Unterschiedsprüfung zu kontrollieren sind. Bisherige Studien hatten mehrfach einen Zusammenhang zwischen Muttersprache/Erstsprache bzw. Migrationshintergrund und SÖS gezeigt. DESI z. B. fand einen deutlichen Zusammenhang zwischen dem SÖS und der Erstsprache (DESI-Konsortium, 2006, S. 22f.). Schüler mit Deutsch als Erstsprache stammten aus Familien mit höherem SÖS als Schüler mit einer anderen Erstsprache als Deutsch. Daher war auch in der vorliegenden Arbeit dieser Zusammenhang zu prüfen, um gegebenenfalls die Einflüsse der Variablen „Schulabschluss der Eltern“ und „Beruf der Eltern“ kontrollieren zu können. Ein deskriptiver Vergleich dieser Variablen zwischen Schülern mit Deutsch als Muttersprache und Schülern mit einer anderen Muttersprache als Deutsch zeigte,

dass tatsächlich die Schüler mit einer anderen Muttersprache als Deutsch im Schnitt aus Familien mit einem niedrigeren SÖS und einem niedrigeren Bildungshintergrund stammen (s. Tab. 57). Um festzustellen, inwiefern dieser Unterschied die Ergebnisse verzerren könnte, wurden daher diese beiden Variablen mit den AVn korreliert. Nur wenn diesbezüglich eine substanzielle Korrelation besteht, ist es notwendig die Variablen zu kontrollieren (vgl. Bortz & Schuster, 2010, S. 311). Auch die Variablen Klassenstufe, Schulart, LRS und Geschlecht wurden mit den AVn korreliert, da diese gegebenenfalls ebenfalls zu kontrollieren waren. Es zeigten sich größtenteils signifikante Korrelationen (s. Tab. 58 in Anhang E).

Bei Berücksichtigung der signifikant mit den Ergebnissen von LESEN 6-7 und LESEN 8-9 korrelierenden Variablen ergaben sich sehr ungleiche Zellbesetzungen, die zudem größtenteils sehr klein waren. Aus diesem Grund wurde ein sogenanntes „Matching“ vorgenommen. Dabei wurde jedem Schüler aus der Gruppe der Schüler mit einer anderen Muttersprache als Deutsch ein Schüler zugeordnet, der diesem hinsichtlich der Merkmale Klassenstufe, Schulart, Geschlecht, LRS-Diagnose, Bildungshintergrund und SÖS möglichst gut entspricht. Dazu wurden die genannten Merkmale entsprechend ihrer Korrelationshöhe mit den AVn absteigend angeordnet und in dieser Reihenfolge bei der Paarbildung berücksichtigt. Auf diese Weise wurden die Merkmale mit dem höchsten Zusammenhang mit den AVn am stärksten berücksichtigt. Anschließend wurden Mittelwertsvergleiche für verbundene Stichproben durchgeführt (vgl. z. B. Bortz & Schuster, 2010, S. 582). Da die Voraussetzungen für *t*-Tests nicht erfüllt und die Stichproben sehr klein waren, wurden nonparametrische Wilcoxon-Tests berechnet und die exakte Auswertung herangezogen.

Der Wilcoxon-Test prüft, ob sich zwei verbundene Stichproben in ihrer zentralen Tendenz signifikant voneinander unterscheiden (Bortz & Schuster, 2010, S. 586). Er eignet sich bei intervallskalierten Daten, die die Voraussetzungen für eine parametrische Auswertung nicht erfüllen. Im Fall von gematchten Stichproben lässt sich jedem Wert aus der einen Gruppe ein entsprechender Wert aus der anderen Gruppe zuordnen, sodass Messwertpaare (sog. „statistische Zwillinge“) entstehen. Die Teststatistik basiert auf einer Bildung einer Rangreihe aus Differenzen der Messwertpaare (Bortz & Schuster, 2010, S. 133f.). Die Absolutbeträge der Differenzen werden in einer Rangreihe angeordnet und Rangplätze vergeben. Die Rangplätze werden nach Vorzeichen der Differenzen getrennt aufsummiert und die Summen verglichen. Je deutlicher sich die Summen voneinander unterscheiden, desto stärker spricht das Ergebnis gegen die Nullhypothese, dass die Gruppen der Stichproben aus Populationen mit gleicher zentraler Tendenz stammen.

Da die Stichproben auch hinsichtlich der Klassenstufe gematcht wurden und der Einfluss der Klassenstufe somit kontrolliert war, konnten hier auch für das Gesamtergebnis von LESEN 6-7 bzw. LESEN 8-9 jeweils beide Klassenstufen gemeinsam ausgewertet werden.

17.3.3 Ergebnisse

Im Folgenden werden zunächst die Ergebnisse der Korrelationsanalysen im Hinblick auf H1a/b (Korrelation mit den Ergebnissen eines anderen Lesetests) und H2a/b (Korrelation mit den Ergebnissen von Lesestrategie-Wissenstests) dargestellt, bevor auf die Ergebnisse der Gruppenvergleiche eingegangen wird. Bei den Gruppenvergleichen werden zuerst die Ergebnisse der Analysen im Hinblick auf H3 (Klassenstufenunterschiede), H4 (Schulartunterschiede) und H5 (Geschlechterunterschiede) dargestellt, anschließend wird auf die Ergebnisse im Hinblick auf H6 (Schüler mit vs. ohne LRS-Diagnose) eingegangen und schließlich werden die Ergebnisse im Hinblick auf H7 (Schüler mit einer anderen Muttersprache als Deutsch vs. Schüler mit Deutsch als Muttersprache) präsentiert.

Hypothesen H1a/b und H2a/b. H1a/b beziehen sich auf die Korrelationen der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit den Ergebnissen des LGVT 6-12. *H1a* besagt, dass die Ergebnisse beider Subtests und das Gesamtergebnis von LESEN 6-7 und LESEN 8-9 substanziell mit den Ergebnissen beider Skalen des LGVT 6-12 korrelieren sollten. Für LESEN 6-7 fallen erwartungskonform die Korrelationen der Ergebnisse des Subtests BLK und des Gesamtergebnisses mit den Ergebnissen beider Skalen des LGVT 6-12 signifikant aus (s. Tab. 23). Die Ergebnisse des Subtests TV von LESEN 6-7 korrelieren mit den Ergebnissen der Skala LGVT-LV ebenfalls erwartungskonform signifikant, die Korrelation mit den Ergebnissen der Skala LGVT-LG fällt hingegen erwartungswidrig nicht signifikant aus. H1a wird somit für LESEN 6-7 größtenteils – jedoch nicht vollständig – bestätigt.

Tabelle 23. Konstruktvalidität: Korrelation der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit Ergebnissen anderer Tests (*N* in Klammern).

	LESEN 6-7			
	BLK	TV	GES 6	GES 7
LGVT 6-12-LG	.62* (452)	.12 (452)	.31* (237)	.64* (215)
LGVT 6-12-LV	.51* (452)	.56* (452)	.61* (237)	.74* (215)
MKT (6. Klasse)	.31* (237)	.34* (237)	.38* (237)	—
WLST 7-12 (7. Klasse)	.33* (215)	.54* (215)	—	.49* (215)
	LESEN 8-9			
	BLK (274)	TV (274)	GES 8 (170)	GES 9 (104)
LGVT 6-12-LG	.72*	.22*	.43*	.65*
LGVT 6-12-LV	.71*	.69*	.83*	.80*
WLST 7-12	.47*	.58*	.74*	.30*

Anmerkungen: *: $p < .01$

Bei LESEN 8-9 fallen die Korrelationen der Subtestergebnisse und des Gesamtergebnisses mit den Ergebnissen beider Skalen des LGVT 6-12 signifikant aus. Dieses Ergeb-

nis bestätigt H1a für LESEN 8-9, wenngleich auch hier die Korrelation der Ergebnisse des Subtests TV von LESEN 8-9 mit den Ergebnissen der Skala LGVT-LG trotz statistischer Signifikanz von geringer Höhe ist.

H1b besagt, dass die Ergebnisse im Subtest BLK von LESEN 6-7 und LESEN 8-9 höher mit den Ergebnissen der Skala LGVT-LG korrelieren sollten als mit den Ergebnissen der Skala LGVT-LV, während die Ergebnisse im Subtest TV von LESEN 6-7 und LESEN 8-9 höher mit den Ergebnissen der Skala LGVT-LV korrelieren sollten als mit den Ergebnissen der Skala LGVT-LG. Für LESEN 6-7 zeigt sich, dass die Ergebnisse des Subtests BLK hoch mit jenen der Skala LGVT-LG und mittelhoch mit jenen der Skala LGVT-LV korrelieren. Der Unterschied der Korrelationshöhen weist die erwartete Richtung auf und fällt signifikant aus bei einer kleinen Effektstärke ($z = 2.43$, $p = .01$, $q = .13$). Die Ergebnisse des Subtests TV von LESEN 6-7 korrelieren nicht signifikant mit den Ergebnissen der Skala LGVT-LG, aber mittelhoch mit den Ergebnissen der Skala LGVT-LV. Die Richtung des Unterschieds der Korrelationshöhen ist dabei ebenfalls erwartungskonform und der Unterschied fällt signifikant aus bei einer großen Effektstärke ($z = -7.68$, $p < .01$, $q = .51$). Die Daten sprechen somit für die Gültigkeit von H1b für LESEN 6-7.

Für LESEN 8-9 zeigt sich, dass die Ergebnisse des Subtests BLK hoch mit den Ergebnissen beider Skalen des LGVT 6-12 korrelieren. Der Unterschied der Korrelationshöhen fällt hier klein und nicht signifikant aus ($z = 0.24$, n.s.). Die Ergebnisse des Subtests TV von LESEN 8-9 korrelieren signifikant – aber in niedriger Höhe – mit den Ergebnissen der Skala LGVT-LG. Zudem korrelieren sie signifikant und erwartungskonform hoch mit den Ergebnissen der Skala LGVT-LV. Der Unterschied der Korrelationshöhen fällt für den Subtest TV von LESEN 8-9 signifikant aus und die Effektstärke ist groß ($z = -7.27$, $p < .01$, $q = .62$). Für LESEN 8-9 kann H1b somit nur teilweise bestätigt werden.

Insgesamt wird H1a für LESEN 6-7 teilweise, für LESEN 8-9 dagegen vollständig bestätigt, während H1b umgekehrt für LESEN 6-7 vollständig und für LESEN 8-9 teilweise bestätigt wird.

H2a/b beziehen sich auf die Korrelationen der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit den Ergebnissen von Lesestrategie-Wissenstests. *H2a* besagt, dass die Subtestergebnisse und die Gesamtergebnisse von LESEN 6-7 substanziell mit den Ergebnissen der Lesestrategie-Wissenstests korrelieren sollten. Bei LESEN 6-7 korrelieren die Ergebnisse beider Subtests und das Gesamtergebnis sowohl mit den Ergebnissen des WLST 7-12 als auch mit jenen des MKT signifikant (s. Tab. 23). Dieses Ergebnis spricht für die Gültigkeit von H2a für LESEN 6-7. Die Korrelationen mit den Ergebnissen des MKT fallen jedoch allesamt relativ niedrig aus, ebenso fällt die Korrelation der Ergebnisse des Subtests BLK von LESEN 6-7 mit den Ergebnissen des WLST 7-12 niedrig aus. Die Ergebnisse des Subtests TV und das Gesamtergebnis von LESEN 6-7 korrelieren dagegen in mittlerer Höhe mit den Ergebnissen des WLST 7-12.

Bei LESEN 8-9 korrelieren die Ergebnisse beider Subtests und das Gesamtergebnis signifikant mit den den Ergebnissen des WLST 7-12, wenngleich die Korrelation des Gesamtergebnisses in der neunten Klassenstufe mit den Ergebnissen des WLST 7-12 relativ niedrig ausfällt. Die Korrelationen der Ergebnisse des WLST 7-12 mit den Subtestergebnissen von LESEN 8-9 fallen beide mittelhoch aus, und die Korrelation mit dem Gesamtergebnis von LESEN 8-9 in der achten Klassenstufe ist hoch. Insgesamt bestätigen diese Ergebnisse somit H2a auch für LESEN 8-9.

H2b besagt, dass die Ergebnisse der Lesestrategie-Wissenstests höher mit den Ergebnissen des Subtests TV als mit den Ergebnissen des Subtests BLK korrelieren sollten. Für LESEN 6-7 korrelieren sowohl die Ergebnisse des MKT als auch die Ergebnisse des WLST 7-12 höher mit den Ergebnissen des Subtests TV als mit den Ergebnissen des Subtests BLK. Beim MKT wird dieser Unterschied zwischen den Korrelationen jedoch nicht signifikant ($z = -0.36$, n.s.), während er beim WLST 7-12 erwartungskonform signifikant ausfällt mit einer kleinen Effektstärke ($z = -2.69$, $p < .01$, $q = .26$). H2b wird somit für LESEN 6-7 nur teilweise bestätigt.

Die Korrelation der Ergebnisse des WLST 7-12 mit den Ergebnissen des Subtests TV von LESEN 8-9 fällt ebenfalls erwartungskonform höher aus als die Korrelation mit den Ergebnissen des Subtests BLK. Der Unterschied ist signifikant mit einer kleinen Effektstärke ($z = -1.77$, $p = .04$, $q = .15$). Dieses Ergebnis bestätigt somit H2b für LESEN 8-9.

Insgesamt wird also H2a für beide Tests bestätigt. H2b hingegen wird für LESEN 6-7 durch die vorliegenden Ergebnisse nur teilweise bestätigt (wobei sich jedoch stets die erwartete Tendenz zeigte), während sie für LESEN 8-9 vollständig bestätigt wird.

Hypothesen H3 bis H5. Im Folgenden werden die Ergebnisse der Prüfung der Hypothesen im Hinblick auf Mittelwertsunterschiede zwischen den Klassenstufen (H3), zwischen den Schularten (H4) und zwischen den Geschlechtern (H5) zunächst für LESEN 6-7 und anschließend für LESEN 8-9 dargestellt (Haupteffekte s. Tab. 24; deskriptive Statistik s. Tab. 53 und 54 in Anhang E).

Multivariat ergibt sich bei LESEN 6-7 ein signifikanter Klassenstufenunterschied mit einer kleinen Effektstärke, ein signifikanter Schulartunterschied mit einer großen Effektstärke und ein signifikanter Geschlechterunterschied mit einer kleinen Effektstärke. Kein Interaktionseffekt erweist sich als signifikant.

Univariat fällt der Klassenstufenunterschied sowohl für den Subtest BLK als auch für den Subtest TV signifikant aus, jeweils mit einer kleinen Effektstärke. Wie erwartet erreichte die siebte Klassenstufe im Mittel jeweils signifikant höhere Ergebniswerte als die sechste Klassenstufe. H3 wird somit für beide Subtests von LESEN 6-7 bestätigt. Auch der Schulartunterschied fällt für beide Subtests signifikant aus, die Effektstärke ist dabei jeweils groß. Der Tamhane-T2-Test zeigt, dass sich alle Schularten bezüglich beider Subtests signifikant voneinander unterscheiden (alle $p < .01$), wobei die Ergebniswerte für die Gymnasiasten am höchsten ausfallen, für die Realschüler im Mittelfeld

liegen und für die Hauptschüler am niedrigsten ausfallen. Auch H4 wird somit für beide Subtests von LESEN 6-7 bestätigt. Der Geschlechterunterschied erweist sich nur für den Subtest BLK als signifikant und zwar mit einer kleinen Effektstärke. Für den Subtest TV zeigt sich kein signifikanter Geschlechterunterschied. H5 bestätigt sich somit für die Subtests von LESEN 6-7 nur teilweise.

Tabelle 24. Haupteffekte der MANOVAs und ANOVAs zur Prüfung von H3 bis H5.

		LESEN 6-7				LESEN 8-9			
Multivariate Ergebnisse bezüglich der Subtests									
UV		F (df1, df2)	p	d	η_p^2	F (df1, df2)	p	d	η_p^2
Klassenstufe		40.74 (2, 1 496)	$< .01$	0.46	.05	10.06 (2, 884)	$< .01$	0.29	.02
Schulart		208.10 (4, 2 944)	$< .01$	1.06	.22	83.39 (4, 1 770)	$< .01$	0.87	.16
Geschlecht		16.34 (2, 1 496)	$< .01$	0.29	.02	15.26 (2, 884)	$< .01$	0.35	.03
Univariate Ergebnisse bezüglich der Subtests									
UV	AV	F (df1, df2)	p	d	η_p^2	F (df1, df2)	p	d	η_p^2
Klassen	BLK	29.73 (1, 1 497)	$< .01$	0.29	.02	< 1			
stufe	TV	66.87 (1, 1 497)	$< .01$	0.41	.04	20.02 (1, 885)	$< .01$	0.29	.02
Schul-	BLK	267.89 (2, 1 497)	$< .01$	1.19	.26	72.38 (2, 885)	$< .01$	0.81	.14
art	TV	423.79 (2, 1 497)	$< .01$	1.50	.36	169.89 (2, 885)	$< .01$	1.28	.28
Ge-	BLK	29.27 (1, 1 497)	$< .01$	0.29	.02	27.67 (1, 885)	$< .01$	0.35	.03
schlecht	TV	< 1				< 1			
Ergebnisse bezüglich des Gesamtergebnisses									
		6. Klasse				8. Klasse			
UV		F (df1, df2)	p	d	η_p^2	F (df1,df2)	p	d	η_p^2
Schulart		316.60 (2, 666)	$< .01$	1.95	.49	131.06 (2, 503)	$< .01$	1.44	.34
Geschlecht		1.74 (1, 666)	n.s.			16.42 (1, 503)	$< .01$	0.35	.03
		7. Klasse				9. Klasse			
UV		F (df1, df2)	p	d	η_p^2	F (df1,df2)	p	d	η_p^2
Schulart		252.87 (2, 831)	$< .01$	1.56	.38	67.50 (2, 382)	$< .01$	1.19	.26
Geschlecht		15.42 (1, 831)	$< .01$	0.27	.02	< 1			

Im Hinblick auf die mittels ANOVA klassenstufenweise geprüften Schulart- und Geschlechterunterschiede bezüglich des Gesamtergebnisses von LESEN 6-7 ergibt sich für die sechste Klassenstufe ein signifikanter Schulartunterschied mit einer großen Effektstärke. Der Scheffé-Post-Hoc-Test zeigt, dass die Unterschiede zwischen allen Schularten signifikant ausfallen (alle $p < .01$), wobei erwartungsgemäß die Ergebniswerte für die Gymnasiasten am höchsten ausfallen, die Ergebniswerte für die Hauptschüler am niedrigsten ausfallen und die Ergebniswerte der Realschüler dazwischen liegen. Der Geschlechterunterschied fällt nicht signifikant aus und auch der Interaktionseffekt nicht. Bezogen auf das Gesamtergebnis wird H4 somit für die sechste Klassenstufe durch die Daten bestätigt, H5 jedoch nicht. Für die siebte Klassenstufe erweisen sich dagegen sowohl der Schulart- als auch der Geschlechterunterschied als signifikant. Der Schulartunterschied erreicht eine hohe Effektstärke. Auch hier unterscheiden sich alle Schularten signifikant voneinander (im Scheffé-Post-Hoc-Test sind

alle $p < .01$) und es zeigt sich das erwartete Muster für die Ergebniswerte (Gymnasium > Realschule > Hauptschule). Im Unterschied zur sechsten Klassenstufe wird für die siebte Klassenstufe auch der Geschlechterunterschied signifikant, der eine kleine Effektstärke aufweist. Der Interaktionseffekt fällt nicht signifikant aus. Für die siebte Klassenstufe werden somit H4 und H5 bestätigt.

Für LESEN 8-9 zeigen sich multivariat ebenfalls ein signifikanter Klassenstufenunterschied mit einer kleinen Effektstärke, ein signifikanter Schulartunterschied mit einer großen Effektstärke sowie ein signifikanter Geschlechterunterschied mit einer kleinen Effektstärke. Bei LESEN 8-9 wird jedoch auch der Interaktionseffekt zwischen Schulart und Geschlecht signifikant, mit einer kleinen Effektstärke ($F(4, 1\,770) = 5.34$, $p < .01$, $d = 0.20$, $\eta_p^2 = .01$). Zwar zeigen die Gymnasiasten bei beiden Geschlechtern für beide Subtests die höchsten Ergebniswerte, die Realschüler liegen im Mittelfeld und die Hauptschüler erreichen die niedrigsten Ergebniswerte, jedoch weist der Geschlechterunterschied nicht in allen Schularten die gleiche Richtung auf. Teilweise erzielen erwartungskonform die Mädchen höhere Ergebniswerte, teilweise erreichen jedoch wider Erwarten die Jungen höhere Ergebniswerte. Weitere Interaktionen fallen nicht signifikant aus.

Univariat fällt der Klassenstufenunterschied nur für den Subtest TV signifikant aus mit einer kleinen Effektstärke. Erwartungskonform erweist sich der Mittelwert der achten Klassenstufe dabei als signifikant niedriger im Vergleich zum Mittelwert der neunten Klassenstufe. Für den Subtest BLK zeigt sich kein signifikanter Klassenstufenunterschied. Somit wird H3 nur für den Subtest TV von LESEN 8-9 bestätigt, nicht jedoch für den Subtest BLK. Der Schulartunterschied fällt für beide Subtests signifikant aus, jeweils mit einer großen Effektstärke. Der Tamhane-T2-Test zeigt, dass die Unterschiede zwischen allen Schularten signifikant sind (alle $p < .01$). Die mittleren Ergebniswerte weisen zudem das erwartete Muster auf (Gymnasium > Realschule > Hauptschule). Der Geschlechterunterschied erweist sich dagegen nur für den Subtest BLK als signifikant mit einer kleinen Effektstärke, nicht aber für den Subtest TV. Aufgrund des signifikanten Interaktionseffektes kann der Geschlechterunterschied allerdings nicht interpretiert werden. Der Interaktionseffekt zwischen Schulart und Geschlecht wird für beide Subtests signifikant, jeweils mit einer kleinen Effektstärke (BLK: $F(2, 885) = 4.18$, $p < .01$, $d = 0.20$, $\eta_p^2 = .01$; TV: $F(2, 885) = 6.02$, $p < .01$. $d = 0.20$, $\eta_p^2 = .01$).

Im Hinblick auf die klassenstufenweise geprüften Unterschiede bezüglich des Gesamtergebnisses von LESEN 8-9 fallen in der achten Klassenstufe sowohl der Schulartunterschied als auch der Geschlechterunterschied signifikant aus. Der Schulartunterschied weist eine große Effektstärke auf. Im Tamhane-T2-Test fallen alle Vergleiche zwischen den Schularten signifikant aus (alle $p < .01$), wobei sich auch hier das erwartete Muster (Gymnasium > Realschule > Hauptschule) zeigt. Der Geschlechterunterschied weist nur eine kleine Effektstärke auf. Die Interaktion der beiden Faktoren fällt nicht signifikant aus. Somit werden für die achte Klassenstufe H4 und H5 bezüglich

des Gesamtergebnisses bestätigt. In der neunten Klassenstufe fällt nur der Schulartunterschied mit einer großen Effektstärke signifikant aus. Der Tamhane-T2-Test zeigt auch hier signifikante Unterschiede zwischen allen Schularten in erwarteter Richtung (Gymnasium $>$ Realschule $>$ Hauptschule; alle $p < .01$). Der Geschlechterunterschied fällt nicht signifikant aus. Somit wird für die neunte Klassenstufe im Hinblick auf das Gesamtergebnis H4 bestätigt, H5 jedoch nicht.

Insgesamt wird also Hypothese H3, die besagt, dass die jeweils höhere Klassenstufe bei beiden Tests besser abschneiden sollte als die jeweils niedrigere, nur teilweise bestätigt. Bei LESEN 6-7 fällt der Klassenstufenunterschied für beide Subtests signifikant aus und weist die erwartete Richtung auf. Bei LESEN 8-9 fällt der Klassenstufenunterschied nur für den Subtest TV signifikant aus. Die Effektstärken sind durchweg als klein zu bezeichnen.

Hypothese H4, welche postuliert, dass bedeutsame Schulartunterschiede bestehen, wird durchweg bestätigt. Es zeigt sich bei beiden Tests für beide Subtests sowie in allen Klassenstufen für das Gesamtergebnis, dass die Gymnasiasten die höchsten Ergebniswerte erzielen, die Realschüler im Mittelfeld liegen und die Hauptschüler die niedrigsten Werte erreichen, wobei die Effektstärke stets groß ausfällt und die Unterschiede jeweils zwischen allen Schularten signifikant sind.

Hypothese H5, laut welcher die Mädchen höhere Ergebniswerte erreichen sollten als die Jungen, wird nur teilweise bestätigt. Bei LESEN 6-7 zeigt sich der erwartete Geschlechterunterschied nur für den Subtest BLK über beide Klassenstufen hinweg sowie für das Gesamtergebnis in der siebten Klassenstufe. Bei LESEN 8-9 kommt es zu einem signifikanten Interaktionseffekt von Geschlecht und Schulart, wobei die Mädchen zum Teil erwartungsgemäß höhere Ergebniswerte erreichen als die Jungen, jedoch auch zum Teil die Jungen wider Erwarten höhere Ergebniswerte erreichen als die Mädchen.

Hypothese H6. Im Folgenden werden die Ergebnisse in Bezug auf H6 dargestellt (Haupteffekte s. Tab. 25, deskriptive Statistik s. Tab. 56 in Anhang E). H6 besagt, dass Schüler mit einer vorliegenden LRS-Diagnose in beiden Subtests sowie im Hinblick auf das Gesamtergebnis von LESEN 6-7 und LESEN 8-9 signifikant niedrigere Ergebniswerte erreichen sollten als Schüler ohne LRS-Diagnose.

Bei LESEN 6-7 ergibt sich bei der multivariaten Prüfung von H6 ein signifikanter Unterschied zwischen Schülern mit LRS und Schülern ohne LRS, wobei die Effektstärke klein ist. Auch der Klassenstufenunterschied erreicht das Signifikanzniveau mit einer ebenfalls kleinen Effektstärke. Der Interaktionseffekt fällt nicht signifikant aus. Univariat zeigen sich für beide Subtests signifikante Unterschiede zwischen Schülern mit LRS und Schülern ohne LRS, wobei die Effektstärke jeweils klein ausfällt. Auch der Klassenstufenunterschied erweist sich für beide Subtests als signifikant, wobei auch hier die Effektstärke jeweils klein ist. Der Interaktionseffekt ist für keinen Subtest signifikant. In Bezug auf das Gesamtergebnis zeigen sich für beide Klassenstufen von

LESEN 6-7 signifikante Unterschiede zwischen Schülern mit LRS und Schülern ohne LRS mit moderaten Effektstärken. Somit bestätigt sich H6 für LESEN 6-7.

Für LESEN 8-9 zeigt sich multivariat ebenfalls ein signifikanter Unterschied zwischen Schülern mit LRS und Schülern ohne LRS. Der Effekt ist allerdings wie bei LESEN 6-7 klein. Der Unterschied zwischen den Klassenstufen erweist sich ebenfalls als signifikant mit einer kleinen Effektstärke. Der Interaktionseffekt ist auch hier nicht signifikant. Univariat ergibt sich ein signifikanter Unterschied zwischen Schülern mit LRS und Schülern ohne LRS für beide Subtests bei einer kleinen Effektstärke. Der Klassenstufenunterschied ist ebenfalls für beide Subtests signifikant mit einer kleinen Effektstärke. Die Interaktionseffekte fallen nicht signifikant aus. Für das Gesamtergebnis erweist sich der Unterschied zwischen Schülern mit LRS und Schülern ohne LRS sowohl für die achte als auch für die neunte Klassenstufe als signifikant. Die Effektstärke ist für die achte Klassenstufe moderat und für die neunte Klassenstufe groß. Auch für LESEN 8-9 wird H6 somit durchweg bestätigt.

Insgesamt bestätigt sich Hypothese H6, welche besagt, dass Schüler mit LRS signifikant niedrigere Ergebniswerte in beiden Tests für beide Subtests und hinsichtlich des Gesamtergebnisses erreichen sollten als Schüler ohne LRS, also sowohl für LESEN 6-7 als auch für LESEN 8-9. Jedoch fallen die Effektstärken für die Subtests durchweg relativ niedrig aus. Für die Gesamtergebnisse sind die Effektstärken hingegen moderat bis groß.

Tabelle 25. Haupteffekte der MANOVAs und Ergebnisse der *t*-Tests zur Prüfung von H6.

		LESEN 6-7				LESEN 8-9			
Multivariate Ergebnisse bezüglich der Subtests									
UV		F (df1, df2)	p	d	η_p^2	F (df1, df2)	p	d	η_p^2
Klassenstufe		6.66 (2, 1 580	$< .01$	0.18	.01	4.30 (2, 923)	.01	0.20	.01
LRS		27.85 (2, 1 580)	$< .01$	0.38	.03	16.41 (2, 923)	$< .01$	0.35	.03
Univariate Ergebnisse bezüglich der Subtests									
UV	AV	F (df1, df2)	p	d	η_p^2	F (df1, df2)	p	d	η_p^2
Klassen-	BLK	6.32 (1, 1 581)	$< .01$	0.13	< 01	7.36 (1, 924)	.01	0.18	.01
stufe	TV	12.00 (1, 1 581)	$< .01$	0.18	.01	23.75 (1, 924)	$< .01$	0.14	.01
LRS	BLK	53.99 (1, 1 581)	$< .01$	0.37	.03	21.43 (1, 924)	$< .01$	0.31	.02
	TV	18.64 (1, 1 581)	$< .01$	0.22	.01	23.75 (1, 924)	$< .01$	0.32	.03
Ergebnisse bezüglich des Gesamtergebnisses									
		6. Klasse				8. Klasse			
UV		t (df)	p	d	η_p^2	t (df)	p	d	η_p^2
LRS		5.19 (712)	$< .01$	0.62	.09	4.60 (524)	$< .01$	0.76	.13
		7. Klasse				9. Klasse			
UV		t (df)	p	d	η_p^2	t (df)	p	d	η_p^2
LRS		5.05 (869)	$< .01$	0.69	.11	3.85 (400)	$< .01$	0.81	.14

Hypothese H7. Bei der Prüfung von H7, welche besagt, dass Schüler mit einer anderen Muttersprache als Deutsch in beiden Subtests sowie hinsichtlich des Gesamtergebnisses sowohl bei LESEN 6-7 als auch bei LESEN 8-9 niedrigere Ergebniswerte erreichen sollten als Schüler mit Deutsch als Muttersprache, wurden Wilcoxon-Tests durchgeführt. Für beide Tests ergaben sich dabei jeweils 17 statistische Zwillinge. Die Ergebnisse können Tabelle 26 entnommen werden.

Tabelle 26. Deskriptive Statistik und Ergebnisse des Wilcoxon-Tests zur Prüfung von H7.

	Deutsch		Andere Sprache		Wilcoxon-Test		
	M	*SD*	*M*	*SD*	*Z*	*p*	*d*
LESEN 6-7							
BLK	54.47	18.42	53.06	12.23	-0.08	n.s.	
TV	20.82	6.63	14.88	5.51	-3.06	$< .01$	0.90
GES	101.12	19.92	92.88	12.15	-1.66	.05	0.40
LESEN 8-9							
BLK	63.24	19.22	59.41	11.07	-1.07	n.s.	
TV	22.29	6.49	17.18	6.28	-2.11	.02	0.62
GES	102.00	18.89	92.82	13.62	-2.06	.02	0.55

Bei LESEN 6-7 fällt das Ergebnis des Wilcoxon-Tests für den Subtest BLK bei einseitiger Testung nicht signifikant aus, für den Subtest TV fällt das Ergebnis dagegen signifikant aus, wobei die Effektstärke groß ist. Für das Gesamtergebnis von LESEN 6-7 liegt das Ergebnis des Wilcoxon-Tests an der Signifikanzgrenze und die Effektstärke ist klein. Auch bei LESEN 8-9 fällt das Ergebnis des Wilcoxon-Tests für den Subtest BLK nicht signifikant aus. Für den Subtest TV und das Gesamtergebnis dagegen erweist es sich als signifikant mit jeweils einer moderaten Effektstärke.

Insgesamt fällt somit der Unterschied zwischen Schülern mit einer anderen Muttersprache als Deutsch und Schülern mit Deutsch als Muttersprache in Bezug auf den Subtest BLK in beiden Tests nicht signifikant aus. Dieses Ergebnis widerspricht H7. Die Ergebnisse in Bezug auf den Subtest TV fallen dagegen für beide Tests signifikant aus. Der Unterschied zwischen den Schülergruppen im Hinblick auf das Gesamtergebnis fällt ebenfalls für beide Tests signifikant aus. Somit wird H7 für den Subtest TV und das Gesamtergebnis sowohl für LESEN 6-7 als auch für LESEN 8-9 bestätigt, für den Subtest BLK dagegen für beide Tests nicht.

17.4 Kriteriumsvalidität

Zur Bestimmung der Kriteriumsvalidität wurden als Außenkriterien die Deutsch- und Mathematiknote sowie der Gesamtnotenschnitt des letzten Schulzeugnisses erhoben. Weiter sollten die Schüler angeben, wie viele Bücher bei ihnen zu Hause vorhanden

sind und wie häufig sie zum Vergnügen lesen. Darüber hinaus wurden die Lehrkräfte gebeten, auf einer Skala von 1 (= sehr schlecht) bis 10 (= sehr gut) die Lesekompetenz jedes einzelnen Schülers zu bewerten.

17.4.1 Hypothesen

Die folgenden Abschnitte fassen wie zuvor bereits bei der Konstruktvalidität zunächst die Forschungsbefunde und theoretischen Grundlagen kurz zusammen, aus welchen sich die Hypothesen zur Bestimmung der Kriteriumsvalidität ergaben. Zuerst wird auf die Hypothesen zur Korrelation der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit dem Skalenwert des Lehrerurteils sowie mit den Schulnoten eingegangen. Anschließend wird die Hypothese zur Korrelation mit der Anzahl der bei den Schülern zu Hause vorhandenen Bücher behandelt, und der letzte Abschnitt beschäftigt sich mit der Hypothese zur Korrelation der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit der Häufigkeit, mit welcher die Schüler zum Vergnügen lesen.

Korrelation mit Lehrerurteil und Schulnoten. Im Zusammenhang mit den Außenkriterien Lehrerurteil, Deutschnote, Mathematiknote und Gesamtnotenschnitt wurde die konvergente und diskriminante Validität von LESEN 6-7 und LESEN 8-9 geprüft. Dabei wurde angenommen, dass die Ergebnisse von LESEN 6-7 und LESEN 8-9 mit einem Kriteriumswert, der ein verwandtes Konstrukt misst, im Sinne konvergenter Validität höher korrelieren sollten als mit einem Kriteriumswert, der ein entfernteres Konstrukt abbildet. Eine niedrige Korrelation der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit einem Kriteriumswert, der ein entferntes Konstrukt misst, kann als Indikator für diskriminante Validität gewertet werden. Als konstruktnah werden das Lehrerurteil bezüglich der Lesekompetenz sowie die Deutschnote angesehen, als konstruktfern dagegen die Mathematiknote. Die Korrelationshöhe mit dem Gesamtnotenschnitt sollte dazwischen liegen. Da sich der Skalenwert des Lehrerurteils rein auf die Lesekompetenz bezieht, sollte dieser zudem höher mit den Ergebnissen von LESEN 6-7 und LESEN 8-9 korrelieren als die Deutschnote, in welche auch andere Aspekte als die Lesekompetenz einfließen.

In bisherigen Studien hat sich allerdings gezeigt, dass der Skalenwert des Lehrerurteils bezüglich der Lesekompetenz in der Sekundarstufe nicht mehr so hoch mit den Ergebnissen standardisierter Lesetests korreliert wie noch in der Grundschule (z. B. Karing, 2009). In der Regel lagen die Korrelationswerte zwischen dem Skalenwert des Lehrerurteils und dem Ergebnis in einem standardisierten Lesetest in der Sekundarstufe im niedrigen Bereich, z. B. bei Karing (2009) im Mittel bei $r = .40$ sowie bei Karing et al. (2011) minderungskorrigiert für eine globale Beurteilung der Lesekompetenz durch die Lehrkraft bei $r = .39$ und für eine aufgabenspezifische Einschätzung der Lesekompetenz durch die Lehrkraft bei $r = .30$. Daher wurde auch für LESEN 6-7 und LESEN 8-9 erwartet, dass die Ergebniswerte zwar mit dem Lehrerurteil höher kor-

relieren als mit den anderen Außenkriterien, aber nicht, dass der Korrelationswert mit dem Lehrerurteil in den Bereich einer hohen Korrelation fällt.

Ferner wurde angenommen, dass eine nicht unerhebliche Korrelation zwischen den Ergebnissen von LESEN 6-7 und LESEN 8-9 mit dem Gesamtnotenschnitt besteht, da Lesen als Schlüsselkompetenz angesehen wird, die nahezu alle Fachbereiche beeinflusst. Diese Korrelationen sollten daher höher sein als die Korrelationen mit der Mathematiknote. Zugleich sollten die Korrelationen der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit dem Gesamtnotenschnitt niedriger sein als die Korrelationen mit der Deutschnote oder dem Lehrerurteil, da angenommen wird, dass LESEN 6-7 und LESEN 8-9 speziell die Lesekompetenz erfassen und nicht die allgemeine Schulleistung. Es ist dabei allerdings zu berücksichtigen, dass sich z. B. bei PISA auch eine hohe Korrelation der Lesekompetenz mit der mathematischen Kompetenz gezeigt hat (Artelt & Schlagmüller, 2004). Die Korrelationen der Ergebnisse des FLVT 5-6 mit Ergebnissen in einem Mathematiktest lagen jedoch im niedrigen Bereich ($r = .26$ bis $r = .27$; Souvignier, Trenk-Hinterberger, Adam-Schwebe & Gold, 2008). Diesbezüglich sind die Vorbefunde also inkonsistent.

Basierend auf dem Ergebnis einer Studie von Harlaar, Kovas, Dale, Petrill und Plomin (2012), die – zumindest für den englischen Sprachraum – zeigte, dass ein engerer Zusammenhang zwischen Leseverständnis und Mathematikleistung besteht als zwischen der reinen Dekodierleistung und der Mathematikleistung, wird weiter angenommen, dass die Korrelation der Ergebnisse des Subtests TV von LESEN 6-7 und LESEN 8-9 mit der Mathematiknote höher ausfallen sollte als die Korrelation der Ergebnisse des Subtests BLK mit der Mathematiknote. Es wurden daher folgende Hypothesen formuliert:

H8a: Die Ergebnisse der Subtests und das Gesamtergebnis von LESEN 6-7 und LESEN 8-9 korrelieren am höchsten mit dem Skalenwert des Lehrerurteils, etwas niedriger mit der Deutschnote, noch etwas niedriger mit dem Gesamtnotenschnitt und am niedrigsten mit der Mathematiknote.

H8b: Die Mathematiknote korreliert sowohl bei LESEN 6-7 als auch bei LESEN 8-9 höher mit dem Ergebnis des Subtests TV als mit dem Ergebnis des Subtests BLK.

Anzahl der Bücher zu Hause. Beispielsweise Chiu und McBride-Chang (2006) fanden einen bedeutsamen positiven Zusammenhang zwischen der Anzahl der Bücher, die bei Schülern zu Hause vorhanden sind, und der Leseleistung. Auch bei PISA 2009 lag in Haushalten, in denen Hörbücher, Bücher mit Gedichten und klassische Literatur verfügbar sind, der Anteil an Schülern auf der untersten Niveaustufe Ia und darunter mit jeweils etwa 8 %, 9 % und 7 % deutlich unter dem deutschen Mittelwert (Naumann et al., 2010). Auch bei der Validierung des LGVT 6-12 lag die Korrelation der Anzahl der Bücher mit den Ergebnissen der Skala LGVT-LG bei $r = .41$ und die mit den Ergebnissen der Skala LGVT-LV bei $r = .25$, wobei beide Korrelationen signifikant ausfielen ($p < .01$, $N = 850$; Schneider et al., 2007). Aus diesem Grund wurde

erwartet, dass auch bei LESEN 6-7 und LESEN 8-9 Schüler mit mehr Büchern zu Hause bessere Ergebniswerte erzielen, was zu folgender Hypothese führte:

H9: Es besteht eine positive Korrelation zwischen den Ergebnissen in beiden Subtests sowie dem Gesamtergebnis von LESEN 6-7 und LESEN 8-9 mit der Anzahl der Bücher, die bei den Schülern zu Hause vorhanden sind.

Lesen zum Vergnügen. Viele Befunde zeigten einen positiven Zusammenhang zwischen Lesefreude und Lesemotivation mit Lesekompetenz (vgl. z. B. Möller & Schiefele, 2004; Paris et al., 1991). Dabei wird vermutet, dass dieser Zusammenhang zum Teil über die Lesemenge vermittelt wird. Bei PISA 2009 liegt die vorhergesagte Lesekompetenz für Schüler, die zum Vergnügen lesen – bei Kontrolle aller anderen Indikatoren von Lesemotivation und Lernstrategien – um etwa 17 Punkte über der Lesekompetenz, die für Schüler erwartet wird, die nicht zum Vergnügen lesen (Naumann et al., 2010). Cipielewski und Stanovich (1992) fanden in ihrer Längsschnittstudie einen Zusammenhang zwischen der Lesehäufigkeit von Schülern und der Entwicklung ihres Leseverständnisses von der dritten bis zur fünften Klasse – selbst wenn die Dekodierleistung auspartialisiert wurde. Daher wird auch für LESEN 6-7 und LESEN 8-9 angenommen, dass die Subtestergebnisse und das Gesamtergebnis in einem positiven Zusammenhang mit der Häufigkeit stehen, mit der Schüler zum Vergnügen lesen. Die entsprechende Hypothese lautet:

H10: Es besteht eine positive Korrelation zwischen der von den Schülern angegebenen Häufigkeit, mit der sie zum Vergnügen lesen, und den Subtestergebnissen sowie dem Gesamtergebnis bei LESEN 6-7 und LESEN 8-9.

17.4.2 Methode

Das Vorgehen erfolgte analog zu den Korrelationsanalysen im Rahmen der Prüfung der Konstruktvalidität. Da für die Außenkriterien jedoch kein Intervalldatenniveau, sondern lediglich Ordinaldatenniveau angenommen wird, wurden zur Prüfung der Kriteriumsvalidität generell Rangkorrelationen (Spearman-Rho) durchgeführt.

Die Korrelationen mit den Schulnoten wurden auf Klassenebene berechnet und anschließend über eine Fisher-*Z*-Transformation unter Berücksichtigung der unterschiedlichen Klassengrößen wie von Bortz und Schuster (2010, S. 160 f.) beschrieben gemittelt. Mit diesem Vorgehen wurde dem Kenntnisstand Rechnung getragen, dass der Maßstab, der bei der Notenvergabe durch die Lehrkraft angelegt wird, von der Leistungsstreuung und dem Leistungsniveau innerhalb der Klassen abhängt und daher zwischen den Klassen sehr unterschiedlich ist (vgl. Kap. 5.2). Aufgrund der Vorgehensweise, dass die Korrelationen mit den Schulnoten auf Klassenebene durchgeführt und anschließend gemittelt wurden, konnten hier auch für das Gesamtergebnis für beide Tests jeweils beide Klassenstufen zusammengefasst werden.

Da es sich um ordinale Daten handelt und zudem für H8a – obwohl es sich um eine abhängige Stichprobe handelt – die Stichprobenumfänge sehr unterschiedlich ausfielen (da die Mathematiknote, der Gesamtnotenschnitt und das Lehrerurteil nur in der Validierungsstichprobe erhoben wurden, die Deutschnote jedoch in der Normstichprobe), wurden die Differenzen zur Prüfung von H8a rein deskriptiv verglichen. Auf eine inferenzstatistische Absicherung wurde verzichtet, da hierfür keine geeignete Formel zur Verfügung stand. Bei H8b variiert die Stichprobengröße nicht, sodass eine inferenzstatistische Auswertung für abhängige Stichproben möglich war.

Aufgrund der Tatsache, dass bei Schulnoten ein numerisch niedriger Wert eine gute Leistung widerspiegelt, während bei LESEN 6-7 bzw. LESEN 8-9 ein numerisch hoher Wert für eine gute Leistung steht, wird für die Schulnoten eine negative Korrelation mit den Testergebnissen erwartet. Die Korrelationen der Testergebnisse mit dem Lehrerurteil sollten dagegen positiv ausfallen, da beim Lehrerurteil ebenso wie bei LESEN 6-7 und LESEN 8-9 ein numerisch hoher Wert für eine gute Leistung spricht. Beim Vergleich der Korrelationen wurden daher schlussendlich stets die Beträge der Korrelationswerte betrachtet.

17.4.3 Ergebnisse

Im Folgenden werden die Ergebnisse der Analysen zur Prüfung der Hypothesen im Zusammenhang mit der Kriteriumsvalidität dargestellt. Es wird erst auf die Hypothesen zum Zusammenhang der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit dem Lehrerurteil und den Schulnoten eingegangen (H8a/b) und anschließend wird der Zusammenhang mit der Anzahl der bei den Schülern zu Hause vorhandenen Bücher (H9) sowie mit der Häufigkeit, mit der Schüler zum Vergnügen lesen (H10), betrachtet.

Hypothesen H8a/b. Erwartungskonform fallen fast alle Korrelationen der Noten mit den Ergebnissen von LESEN 6-7 negativ aus (s. Tab. 27). Alle Korrelationen bis auf jene mit der Mathematiknote sind auf dem 1%-Niveau signifikant. Zudem müssen alle Korrelationen erwartungskonform als eher niedrig bezeichnet werden. Lediglich die Korrelation des Lehrerurteils mit dem Gesamtergebnis von LESEN 6-7 kann als mittelhoch bezeichnet werden. Dennoch zeigt sich insgesamt das aufgrund von *H8a* erwartete Muster für beide Subtests und das Gesamtergebnis. Betrachtet man die Beträge der Korrelationswerte, erweist sich der Zusammenhang der Subtestergebnisse von LESEN 6-7 mit dem Lehrerurteil jeweils als am höchsten, gefolgt von einem etwas niedrigeren Zusammenhang mit der Deutschnote, einem nochmals etwas niedrigeren Zusammenhang mit dem Gesamtnotenschnitt und dem niedrigsten Zusammenhang mit der Mathematiknote. Die Daten bestätigen somit H8a für LESEN 6-7.

Bei LESEN 8-9 zeigt sich bei Betrachtung der Beträge der Korrelationen für die Ergebnisse des Subtests TV und das Gesamtergebnis ebenfalls das erwartete Muster mit der höchsten Korrelation mit dem Skalenwert des Lehrerurteils, einer etwas niedri-

geren Korrelation mit der Deutschnote, einer noch niedrigeren Korrelation mit dem Gesamtnotenschnitt und der niedrigsten Korrelation mit der Mathematiknote (s. Tab. 27). Beim Subtest BLK fällt die Korrelation der Ergebnisse mit dem Skalenwert des Lehrerurteils aus dem Rahmen. Diese fällt erwartungswidrig sehr gering und nicht signifikant aus. In Bezug auf die Noten zeigt sich jedoch auch für den Subtest BLK das erwartete Muster. Mit Ausnahme der Korrelationen der Ergebnisse des Subtests BLK mit dem Skalenwert des Lehrerurteils und der Mathematiknote fallen alle Korrelationen auf dem 1%-Niveau signifikant aus. H8a kann für LESEN 8-9 also nur teilweise bestätigt werden.

Tabelle 27. Kriteriumsvalidität: Korrelation (Spearman-Rho) mit Außenkriterien.

	LESEN 6-7				LESEN 8-9			
	N	BLK	TV	GES	*N*	BLK	TV	GES
Lehrerurteil	286	.33**	.38**	.49**	170	.14	.58**	.44**
Deutschnote	1 512	-.28**	-.25**	−.31**	852	-.26**	-.35**	-.39**
Notenschnitt	364	-.17**	-.20**	-.25**	184	-.26**	-.27**	-.33**
Mathematiknote	362	.02	-.13*	-.12*	191	-.15	-.22**	-.22**

Anmerkungen: *: $p < .05$; **: $p < .01$

Wie mit *H8b* erwartet, fällt bei LESEN 6-7 die Korrelation der Mathematiknote mit den Ergebnissen im Subtest TV höher aus als die mit den Ergebnissen im Subtest BLK. Der Unterschied ist signifikant und weist eine kleine Effektstärke auf ($z = -1.85$; $p = .03$; $q = .10$). Dieses Ergebnis bestätigt also H8b für LESEN 6-7. Auch bei LESEN 8-9 zeigt sich erwartungskonform, dass die Ergebnisse des Subtests TV höher mit der Mathematiknote korrelieren als die Ergebnisse des Subtests BLK, wenngleich der Unterschied nicht signifikant ausfällt ($z = 1.06$; n.s.). H8b kann für LESEN 8-9 somit nur tendenziell bestätigt werden.

Insgesamt wird H8a für LESEN 6-7 vollständig und für LESEN 8-9 nur teilweise bestätigt. In Bezug auf H8b kann das erwartungskonforme Ergebnis für LESEN 6-7 zufallskritisch abgesichert werden, während sich für LESEN 8-9 zwar die erwartete Tendenz zeigt, jedoch das Signifikanzniveau verfehlt wird.

Hypothese H9. Die Korrelationen zwischen den Subtestergebnissen und dem Gesamtergebnis von LESEN 6-7 und LESEN 8-9 mit der Angabe der Schüler zur Anzahl der Bücher, die bei ihnen zu Hause vorhanden ist, fallen erwartungskonform durchgehend signifikant aus. Bei LESEN 6-7 ist die Korrelation für den Subtest BLK mit einem Wert von $r = .36$ als relativ niedrig zu bezeichnen, für den Subtest TV liegt der Wert mit $r = .44$ im mittelhohen Bereich ($p < .01$; $N = 449$ für beide Subtests). Die Korrelation mit dem Gesamtergebnis von LESEN 6-7 liegt für die sechste Klassenstufe bei $r = .45$ ($p < .01$, $N = 198$), für die siebte Klassenstufe bei $r = .49$ ($p < .01$, $N = 186$) und somit für beide Klassenstufen im ebenfalls mittelhohen Bereich.

Bei LESEN 8-9 zeigt sich ein ähnliches Bild: Die Korrelation der Anzahl der bei den Schülern zu Hause vorhandenen Bücher mit den Ergebnissen des Subtests BLK fällt mit $r = .37$ relativ niedrig aus, während sie für die Ergebnisse des Subtests TV mit $r = .52$ in den mittelhohen Bereich fällt (für beide Subtests: $p < .01$, $N = 270$). Für das Gesamtergebnis von LESEN 8-9 fallen die Korrelationen mit einem Wert von $r = .50$ ($p < .01$, $N = 149$) für die achte Klassenstufe und einem Wert von $r = .46$ ($p < .01$, $N = 100$) für die neunte Klassenstufe ebenfalls wieder mittelhoch aus. H9 wird somit für beide Tests bestätigt.

Hypothese H10. Die Angabe der Schüler zur Häufigkeit, mit der sie zum Vergnügen lesen, korreliert bei LESEN 6-7 und LESEN 8-9 signifikant mit beiden Subtestergebnissen. Bei LESEN 6-7 fallen die Korrelationen mit Werten von $r = .20$ für den Subtest BLK und $r = .23$ für den Subtest TV jedoch relativ niedrig aus ($p < .01$; $N = 452$ für beide Subtests). Auch die Korrelationen mit dem Gesamtergebnis von LESEN 6-7 sind als niedrig zu bezeichnen, sie liegen bei $r = .24$ ($p < .01$, $N = 199$) für die sechste Klassenstufe und bei $r = .26$ ($p < .01$, $N = 188$) für die siebte Klassenstufe.

Bei LESEN 8-9 zeigen sich etwas höhere Werte. Mit $r = .39$ fällt die Korrelation für den Subtest BLK zwar ebenfalls eher niedrig aus, die Korrelation für den Subtest TV erweist sich jedoch mit $r = .45$ als mittelhoch ($p < .01$; $N = 274$). Die Korrelationen mit dem Gesamtergebnis fallen mit $r = .52$ ($p < .01$, $N = 152$) für die achte Klassenstufe und $r = .51$ ($p < .01$, $N = 101$) für die neunte Klassenstufe ebenfalls in einen Bereich mittlerer Höhe. H10 wird also durch die vorliegenden Daten für beide Tests bestätigt.

17.5 Diskussion

Im Rahmen der Validitätsanalysen für LESEN 6-7 und LESEN 8-9 wurden die gängigen Validitätsaspekte Inhaltsvalidität, Konstruktvalidität und Kriteriumsvalidität betrachtet. Im Folgenden werden die entsprechenden Ergebnisse zusammengefasst, diskutiert und mit Vorbefunden verglichen. Außerdem wird das methodische Vorgehen kritisch beleuchtet.

Inhaltsvalidität. Die Inhaltsvalidität von LESEN 6-7 und LESEN 8-9 lässt sich theoretisch und logisch begründen. Die Tests erfassen umfassend die relevantesten Aspekte des Leseverständnisses. Dabei ist jedoch anzumerken, dass für den Subtest TV zwar zunächst theoriegeleitet gleich viele Items für jede Verständnisebene erstellt wurden und somit eine sehr hohe inhaltliche Validität angestrebt wurde. Die Items der verschiedenen Ebenen konnten sich aber nur in unterschiedlichem Maße in der empirischen Erprobung bewähren, weshalb es letztendlich nicht gelang, die gleichmäßige Verteilung beizubehalten. In den Endversionen der Subtests TV sind Items, die sich auf die Metaebenen beziehen, deutlich unterrepräsentiert. Da diese Items bei der Itemse-

lektion zum Großteil eliminiert werden mussten, wäre zu überlegen, ob es sich dabei um eine distinkte Skala handeln könnte, die separat erfasst werden sollte. Inhaltlich sind die Metaebenen für die Sekundarstufe auf jeden Fall interessant, denn Schüler dieser Klassenstufen sollten beispielsweise lernen, Texte richtig zu interpretieren und die Intention des Autors zu verstehen. Insbesondere beim eigenständigen Reflektieren und Bewerten zeigten jedoch die deutschen Schüler bei PISA 2009 eine relative Schwäche (Naumann et al., 2010).

Konstruktvalidität. Zur Prüfung der Konstruktvalidität wurden die Ergebnisse von LESEN 6-7 und LESEN 8-9 mit den Ergebnissen anderer standardisierter Tests korreliert. Zudem wurde geprüft, ob LESEN 6-7 und LESEN 8-9 empirische Vorbefunde replizieren können und in Bezug auf aus der Theorie abgeleitete Hypothesen zu erwartungskonformen Ergebnissen kommen.

H1a besagte, dass die Ergebnisse beider Subtests und das Gesamtergebnis von LESEN 6-7 und LESEN 8-9 substanziell mit den Ergebnissen beider Skalen des LGVT 6-12 korrelieren sollten. H1b postulierte weiter, dass die Ergebnisse des Subtests BLK von LESEN 6-7 und LESEN 8-9 mit den Ergebnissen der Skala LGVT-LG höher korrelieren sollten als mit den Ergebnissen der Skala LGVT-LV, und dass umgekehrt die Ergebnisse des Subtests TV höher mit den Ergebnissen der Skala LGVT-LV korrelieren sollten als mit den Ergebnissen der Skala LGVT-LG. H1a konnte für LESEN 6-7 nur teilweise bestätigt werden, für LESEN 8-9 jedoch vollständig. H1b konnte umgekehrt für LESEN 6-7 vollständig und für LESEN 8-9 nur teilweise bestätigt werden.

Die Abweichungen einzelner Korrelationen vom erwarteten Muster im Hinblick auf H1a und H1b könnten zum einen auf tatsächliche Mängel von LESEN 6-7 und LESEN 8-9 zurückzuführen sein, zum anderen könnten sie aber auch mit Mängeln des LGVT 6-12 zusammenhängen. Beispielsweise werden beim LGVT 6-12 die Lesegeschwindigkeit und das Leseverständnis stark konfundiert erfasst – stärker konfundiert als bei LESEN 6-7 und LESEN 8-9. Darüber hinaus gibt es bereits empirische Hinweise auf die eingeschränkte Validität des LGVT 6-12, denn dessen Ergebnisse korrelieren nur niedrig mit dem Skalenwert des Lehrerurteils und den Werten anderer Außenkriterien (W. Lenhard, 2013, S. 99). Auch die Korrelation mit den Ergebnisse anderer Tests (z. B. Lesefortschrittsdiagnostik, LDL) fallen niedrig aus (z. B. W. Lenhard, 2013, S. 97). Die Ergebnisse der Skala LGVT-LG korrelieren zudem nur niedrig mit dem PISA-Lesetest (Schneider, 2009).

Der Einsatz alternativer Gruppentests zur Prüfung der Konstruktvalidität von LESEN 6-7 und LESEN 8-9 wäre jedoch für die basale Lesekompetenz nur bis zur achten Klassenstufe anhand des SLS 5-8 möglich gewesen und für das Leseverständnis nur in der sechsten Klassenstufe anhand des FLVT 5-6. Für die basale Lesekompetenz ab der neunten Klassenstufe und das Leseverständnis ab der siebten Klassenstufe stand zum Zeitpunkt der Datenerhebung für die Validierung von LESEN 6-7 und LESEN 8-9 keine Alternative zum LGVT 6-12 zur Verfügung. Trotz der beschriebenen Kritikpunkte am

LGVT 6-12 sprechen die größtenteils erwartungskonform ausfallenden Ergebnisse in Bezug auf H1a/b jedoch für die Konstruktvalidität von LESEN 6-7 und LESEN 8-9.

H2a besagte, dass die Ergebnisse beider Subtests und das Gesamtergebnis von LESEN 6-7 und LESEN 8-9 substanziell mit den Ergebnissen der eingesetzten Lesestrategie-Wissenstests korrelieren sollten. Diese Erwartung erfüllte sich für beide Tests. H2b, welcher zufolge die Ergebnisse der Lesestrategie-Wissenstests mit den Ergebnissen des Subtests TV höher korrelieren sollten als mit den Ergebnissen des Subtests BLK, wurde für LESEN 8-9 bestätigt. Bei LESEN 6-7 dagegen fällt der Unterschied nur für den WLST 7-12 signifikant aus, während das Ergebnis für den MKT zwar die entsprechende Tendenz aufweist, aber das Signifikanzniveau verfehlt. Letzteres könnte damit zusammenhängen, dass bei LESEN 6-7 die Stichproben aufgrund der Separierung nach Klassenstufen kleiner ausfielen und somit die Wahrscheinlichkeit für ein signifikantes Ergebnis geringer war. Andererseits ist augenscheinlich, dass die Korrelationshöhe der Ergebnisse des MKT für die sechste Klassenstufe mit den Ergebnissen des Subtests BLK von LESEN 6-7 nur geringfügig niedriger ist als die Korrelationshöhe mit den Ergebnissen des Subtests TV und das Verfehlen des Signifikanzniveaus somit nicht nur auf die geringe Stichprobengröße zurückzuführen sein dürfte. Insgesamt stimmen die Ergebnisse in Bezug auf den Zusammenhang zwischen Lesestrategiewissen und Leseleistung, insbesondere Lese*verständnis*leistung, mit vorherigen Forschungsergebnissen überein (vgl. z. B. D. H. Rost & Buch, 2010), was für die Konstruktvalidität von LESEN 6-7 und LESEN 8-9 spricht.

Laut H3 sollten sowohl bei LESEN 6-7 als auch bei LESEN 8-9 die Schüler der jeweils höheren Klassenstufe (7 bzw. 9) höhere Ergebniswerte erreichen als die Schüler der jeweils niedrigeren Klassenstufe (6 bzw. 8). Diese Hypothese wurde für LESEN 6-7 komplett und für LESEN 8-9 teilweise bestätigt. Bei LESEN 8-9 fällt der Klassenstufenunterschied nur für den Subtest TV signifikant aus. Diese Ergebnisse fügen sich gut in die bisherige Befundlage ein. Da die Leseleistung in der Sekundarstufe nicht mehr in dem Ausmaß explizit gefördert wird wie in der Grundschule und basale Lesekompetenzen bei dem größten Teil der Schüler als gegeben angesehen werden können (vgl. Naumann et al., 2010), wurden zwar Unterschiede zwischen den Klassenstufen erwartet, jedoch wurde zugleich angenommen, dass die Unterschiede nicht mehr so groß ausfallen wie in der Grundschule. DESI z. B. fand im Verlauf des neunten Schuljahres keinen generellen signifikanten Leistungszuwachs im Lesen, nur im Gymnasium stieg die Leistung signifikant an (DESI-Konsortium, 2006, S. 6f.). Auch Hulslander, Olson, Willcutt und Wadsworth (2010) fanden beim Leseverständnis zwar im Mittel Anstiege im Verlauf der Sekundarschulzeit, jedoch fielen diese nicht für alle Schüler groß aus. Insgesamt zeigten sich in den ersten Sekundarschuljahren in früheren Studien Leistungszuwächse von etwa einer drittel bis einer halben Standardabweichung pro Schuljahr (z. B. Retelsdorf & Möller, 2008; Baumert, 2002; R. Lehmann & Lenkeit, 2008). Damit stimmen die im Rahmen der vorliegenden Arbeit gefundenen Ergebnisse gut überein. Der Klassenstufenunterschied bei LESEN 6-7 beträgt multivariat 0.46

Standardabweichungen sowie univariat 0.29 Standardabweichungen für den Subtest BLK und 0.41 Standardabweichungen für den Subtest TV. Bei LESEN 8-9 beträgt der Klassenstufenunterschied multivariat 0.29 Standardabweichungen sowie univariat für den Subtest TV ebenfalls 0.29 Standardabweichungen. Für den Subtest BLK fällt der Unterschied bei LESEN 8-9 nicht signifikant aus. Dass die Leistungsunterschiede zwischen den Klassenstufen für LESEN 8-9 kleiner ausfallen als für LESEN 6-7, passt ebenfalls zu bisherigen Befunden, die zeigten, dass die Leistungszuwächse im Verlauf der Sekundarstufe immer stärker abflachen (W. Lenhard, 2013; Philipp, 2011b).

H4 zufolge sollten sowohl bei LESEN 6-7 als auch bei LESEN 8-9 in beiden Subtests sowie auch in Bezug auf das Gesamtergebnis die Gymnasiasten die höchsten Ergebniswerte erzielen, gefolgt von den Realschülern und die niedrigsten Ergebniswerte wurden von den Hauptschülern erwartet. H4 wurde bei beiden Tests für beide Subtests sowie für das Gesamtergebnis bestätigt. Die Effektstärken sind jeweils groß, und die Post-Hoc-Tests zeigen, dass die Unterschiede zwischen allen Schularten signifikant sind. Der Varianzanteil, der durch die Schulart erklärt wird, liegt bei LESEN 6-7 zwischen 22 % und 49 % und bei LESEN 8-9 zwischen 14 % und 28 %. Damit liegen die Werte größtenteils unter den Ergebnissen, die von PISA 2000 und DESI für die Varianzaufklärung durch die Schulart berichtet werden. Bei PISA 2000 konnten 45 % der Varianz in der Lesekompetenz auf die Schulart zurückgeführt werden und bei DESI 52 % im Bereich Deutsch (DESI-Konsortium, 2006, S. 54). Zu beachten ist bei der Interpretation der Ergebnisse, dass die Schulartzugehörigkeit generell mit verschiedenen anderen Variablen konfundiert ist (z. B. Schul- und Klassenzugehörigkeit oder Intelligenz) und die genannten Ergebnisse somit keine Kausalinterpretationen erlauben. Dass es keine signifikante Interaktion von Klassenstufe und Schulart gibt, passt außerdem gut zu den Ergebnissen bisheriger Studien, die im Längsschnitt zeigten, dass in der Sekundarstufe ein Leistungszuwachs stattfindet, der für alle Leistungsniveaus parallel verläuft (vgl. W. Lenhard, 2013, S. 42). Die großen Effektstärken bezüglich der Schulartunterschiede machen darüber hinaus deutlich, dass die Nichtrepräsentativität der Normstichprobe hinsichtlich der Verteilung der Schüler über die verschiedenen Schularten kritisch zu sehen ist und die Gültigkeit der schulartübergreifenden Normen einschränkt.

H5, welche besagt, dass die Mädchen in beiden Subtests und im Gesamtergebnis sowohl bei LESEN 6-7 als auch bei LESEN 8-9 höhere Ergebniswerte erreichen sollten als die Jungen, wurde für beide Tests nur teilweise bestätigt. Bei LESEN 6-7 zeigt sich ein signifikanter Geschlechterunterschied nur für den Subtest BLK und für das Gesamtergebnis in der siebten Klassenstufe. Für den Subtest TV fällt der Unterschied dagegen nicht signifikant aus, ebensowenig für das Gesamtergebnis in der sechsten Klassenstufe. Dort, wo der Unterschied das Signifikanzniveau erreicht, ist zudem die Effektstärke klein. Bei LESEN 8-9 zeigt sich für beide Subtests ein Interaktionseffekt von Geschlecht und Schulart, wobei der Geschlechterunterschied nicht in allen Schularten die gleiche Richtung aufweist. Teilweise erreichten die Jungen erwartungswidrig

höhere Ergebniswerte als die Mädchen. Diese Ergebnisse widersprechen insofern den Erwartungen, als bisherige Studien, z. B. PISA (Naumann et al., 2010), kontinuierlich und in allen Bereichen der Lesekompetenz Geschlechterunterschiede zugunsten der Mädchen fanden. Die erwartungswidrigen Ergebnisse könnten somit einerseits gegen die Validität von LESEN 6-7 und LESEN 8-9 sprechen. Andererseits fallen die Geschlechterunterschiede aber insgesamt in nahezu allen Studien, in denen sie auftreten, in einen Bereich von unbedeutender bis moderater Effektstärke, und einzelne Studien finden auch keine signifikanten Geschlechterunterschiede (vgl. z. B. W. Lenhard, 2013; S. Meyer, 2009). Da Geschlechterunterschiede in der Lesekompetenz außerdem häufig auf den Inhalt von Lesestoffen zurückgeführt werden, könnte man spekulieren, dass das Ausbleiben der erwarteten Geschlechterunterschiede in der vorliegenden Arbeit in der Themenwahl bei den Texten für die Subtests TV begründet liegt. Somit könnte hierin sogar ein Vorteil von LESEN 6-7 und LESEN 8-9 liegen.

Bezüglich H6 fällt das Ergebnismuster bei beiden Tests sowohl für beide Subtests als auch für das Gesamtergebnis erwartungskonform aus. Schüler mit einer LRS-Diagnose erreichten jeweils signifikant niedrigere Ergebniswerte als Schüler ohne eine LRS-Diagnose. Für beide Subtests fallen die Effekte allerdings klein aus, für das Gesamtergebnis dagegen zeigen sich moderate bis große Effekte. Da sich LRS in der Sekundarstufe vor allem in Problemen bei der Rechtschreibung äußert, die Lesekompetenz in den mittleren und höheren Klassenstufen dagegen im Vergleich zur Grundschulleistung meist deutlich verbessert ist, entsprechen die kleinen Effektstärken den Erwartungen. Dass sich im Hinblick auf das Gesamtergebnis ein größerer Unterschied zwischen Schülern mit und ohne LRS zeigt, könnte mit einer höheren Reliabilität des Gesamtergebnisses zusammenhängen.

Mit H7 wurde postuliert, dass Schüler mit einer anderen Muttersprache als Deutsch in beiden Subtests sowie hinsichtlich des Gesamtergebnisses von LESEN 6-7 und LESEN 8-9 niedrigere Ergebniswerte erzielen sollten als Schüler mit Deutsch als Muttersprache. Bei der Auswertung wurde durch ein „Matching“ der Einfluss zahlreicher Variablen (Klassenstufe, Schulart, Geschlecht, Beruf der Eltern, Schulabschluss der Eltern) kontrolliert. H7 bestätigte sich bei beiden Tests nur für den Subtest TV und das Gesamtergebnis – und dies jeweils mit moderater bis großer Effektstärke. Im Subtest BLK wurden keine signifikanten Unterschiede zwischen den Schülergruppen gefunden. Das Ergebnis in Bezug auf den Subtest BLK widerspricht zwar H7, passt jedoch mit Befunden von Limbird und Stanat (2006) sowie A. E. Marx und Stanat (2011) zusammen, die zeigten, dass in Bezug auf die Leseflüssigkeit auf Wortebene sowie zum Teil auch beim Lesen von kurzen – aus zwei bis drei Sätzen bestehenden – Texten keine bedeutsamen Unterschiede zwischen Schülern mit Migrationshintergrund und Schülern ohne Migrationshintergrund bestehen. In diesen Studien ergaben sich – wie in der vorliegenden Arbeit – lediglich im Hinblick auf das Textverständnis Unterschiede. Auch bei der Validierung des ZLT-II (Petermann & Daseking, 2012) zeigten sich bei Aufgaben, die basale Aspekte der Lesekompetenz (Lesegeschwindigkeit und -genauigkeit

beim Lesen von Einzellauten, Lautverbindungen, Wörtern und kurzen Textabschnitten) erfassen, keine Unterschiede zwischen Schülern mit Deutsch als Muttersprache und Schülern mit einer anderen Muttersprache als Deutsch. Zwar ist beim Subtest BLK von LESEN 6-7 und LESEN 8-9 ein Mindestmaß an Leseverständnis nötig, um die Sätze korrekt hinsichtlich ihrer inhaltlichen Richtigkeit beurteilen zu können, dennoch liegt der Schwerpunkt auf der Lesegeschwindigkeit, weshalb ein Vergleich der Ergebnisse mit jenen von Limbird und Stanat (2006), A. E. Marx und Stanat (2011) und (Petermann & Daseking, 2012) plausibel erscheint.

Bei der Interpretation der Ergebnisse der vorliegenden Arbeit bezüglich H7 ist jedoch zu berücksichtigen, dass erstens die Entscheidung darüber, welcher Kategorie ein Beruf zugeordnet werden sollte, häufig nicht eindeutig getroffen werden konnte, dass zweitens von vielen Schülern die nötigen Angaben zu den Berufen der Eltern fehlten und dass drittens bei den vorhandenen Angaben teilweise die Zuverlässigkeit angezweifelt werden muss. Viele Schüler gaben zudem an, nicht zu wissen, welchen Beruf ihre Eltern ausüben oder nannten den Namen der Firma, in der ihr Vater/ihre Mutter arbeitet, anstatt die Berufsbezeichnung. Insbesondere niedrigere Schulabschlüsse und „einfachere“ Berufe scheinen in den verfügbaren Daten unterrepräsentiert zu sein. In Bezug auf den Bildungshintergrund erwiesen sich insbesondere die Angaben von Schülern mit Migrationshintergrund als nicht verwertbar. Vor allem Schüler, deren Eltern einen Schulabschluss im Ausland gemacht haben, konnten häufig nicht angeben, welcher Art dieser Schulabschluss war, sie gaben z. B. lediglich an „nicht in Deutschland“ oder „Schulabschluss in der Türkei“. Dadurch wurde der Anteil an Kindern mit Migrationshintergrund in der Auswertung stark verringert. Insgesamt ergaben sich für LESEN 6-7 und LESEN 8-9 jeweils nur 17 „statistische Zwillinge“ für die Auswertung. Die Ergebnisse bezüglich H7 sind daher trotz Berücksichtigung zahlreicher Kontrollvariablen mit Vorsicht zu interpretieren.

Insgesamt fallen die Befunde zur Beurteilung der Konstruktvalidität größtenteils erwartungskonform aus und zeigen, dass LESEN 6-7 und LESEN 8-9 Vorbefunde replizieren und die Testergebnisse auch im Hinblick auf aus der Theorie abgeleitete Hypothesen erwartungskonform ausfallen. Die Ergebnisse sprechen somit für die Konstruktvalidität von LESEN 6-7 und LESEN 8-9.

Kriteriumsvalidität. Zur Prüfung der Kriteriumsvalidität wurden Hypothesen darüber aufgestellt, wie hoch die Ergebnisse von LESEN 6-7 und LESEN 8-9 mit verschiedenen Außenkriterien (Lehrerurteil zur Lesekompetenz, Schulnoten, Anzahl der zu Hause vorhandenen Bücher, Häufigkeit des Lesen zum Vergnügen) korrelieren sollten.

H8a besagte, dass die Ergebnisse der Subtests und das Gesamtergebnis von LESEN 6-7 und LESEN 8-9 am höchsten mit dem Skalenwert des Lehrerurteils bezüglich der Lesekompetenz, etwas niedriger mit der Deutschnote, noch etwas niedriger mit dem Gesamtnotenschnitt und am niedrigsten mit der Mathematiknote korrelieren sollten. Dieses erwartete Muster wird für LESEN 6-7 uneingeschränkt und für LESEN 8-9 mit

einer Ausnahme (Korrelation des Skalenwertes des Lehrerurteils mit dem Ergebnis im Subtest BLK, welche niedriger ausfiel als erwartet) bestätigt. Dieses größtenteils erwartungskonforme Ergebnis zeigt, dass LESEN 6-7 und LESEN 8-9 nicht nur die allgemeine (Schul-)Leistungsfähigkeit, sondern spezifisch eine bestimmte Kompetenz erfassen. Insgesamt fallen jedoch selbst die Korrelationen mit dem Skalenwert des Lehrerurteils nur niedrig bis mittelhoch aus. Bis auf die Korrelation der Ergebnisse des Subtests BLK von LESEN 8-9 mit dem Skalenwert des Lehrerurteils, die nicht signifikant ausfällt, weisen alle anderen Korrelationen mit Werten von $r = .33$ bis $r = .58$ allerdings mindestens die Höhe der von anderen Studien berichteten Korrelationswerte für die Sekundarstufe auf, die für die Korrelation eines Skalenwertes des Lehrerurteils mit den Ergebnissen in standardisierten Lesetests zwischen $r = .30$ und $r = .40$ liegen (vgl. Karing, 2009; Karing et al., 2011).

Interessanterweise äußerten zahlreiche Lehrkräfte, dass sie die Lesekompetenz ihrer Schüler nur schwer einschätzen können. Einige weigerten sich sogar, es überhaupt zu versuchen. Dies könnte daran liegen, dass in der Sekundarstufe die Lesekompetenz nicht mehr explizit geprüft wird (vgl. auch Kap. 5.2), was auch die relativ niedrigen Korrelationen mit der Deutschnote erklären könnte. In der Sekundarstufe fließt die Lesekompetenz nur noch indirekt in die Deutschnote ein. Zudem ist anzumerken, dass die Höhe der Korrelationen der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit der Deutschnote und dem Skalenwert des Lehrerurteils zwischen den Klassen stark variiert. Dies stimmt mit Vorbefunden überein, die zeigten, dass die diagnostische Kompetenz von Lehrkräften interindividuell sehr unterschiedlich ausgeprägt ist (Artelt, 2009). Die Höhe der Korrelationen der Ergebniswerte von LESEN 6-7 und LESEN 8-9 mit der Deutschnote ($r = -.25$ bis $r = -.39$) stimmt gut mit den in den Testmanualen anderer standardisierter Lesetests berichteten Werte überein (vgl. auch 5.3.4). Für den ZLT-II werden z. B. in der Sekundarstufe geringere Korrelationen der Lesetestergebnisse mit den Schulnoten angegeben als für die Grundschule (Petermann & Daseking, 2012). In der weiterführenden Schule korrelierten beim ZLT-II die Schulnoten mit der Lesezeit beim Lesen von Textabschnitten zu $r = -.31$ und mit der entsprechenden Fehlerzahl zu $r = .36$. Ähnlich fallen auch bei der Validierung des FLVT 5-6 die Korrelationen mit der Deutschnote aus (Form A: $r = -.38$; Form B: $r = -.32$; Souvignier et al., 2008). In Bezug auf die Mathematiknote zeigt sich eine etwas geringere Korrelation mit den Ergebniswerten von LESEN 6-7 und LESEN 8-9 im Vergleich zu den für den FLVT 5-6 berichteten Korrelationswerten (Souvignier et al., 2008), was für die diskriminante Validität von LESEN 6-7 und LESEN 8-9 spricht.

H8b postulierte darüber hinaus, dass die Mathematiknote stärker mit den Ergebnissen im Subtest TV von LESEN 6-7 und LESEN 8-9 korrelieren sollte als mit den Ergebnissen im Subtest BLK. Diese Hypothese wird für LESEN 6-7 bestätigt und zufallskritisch abgesichert, während sie für LESEN 8-9 nur tendenziell bestätigt wird. Bei der Interpretation dieses Ergebnisses ist zu beachten, dass bei LESEN 8-9 die Stichprobe deutlich kleiner war als bei LESEN 6-7, weshalb die Wahrscheinlichkeit für ein signifi-

kantes Ergebnis geringer war. Somit stimmen die Ergebnisse in Bezug auf LESEN 6-7 und LESEN 8-9 für die deutsche Validierungsstichprobe mit dem Befund von Harlaar et al. (2012) aus dem englischsprachigen Raum überein, der zeigte, dass die Mathematikleistung stärker mit dem Leseverständnis zusammenhängt als mit der Dekodierfähigkeit, wobei sich der Befund in der vorliegenden Arbeit jedoch nur für die sechste und siebte Klasse zufallskritisch absichern ließ.

Mit der durchgehenden Bestätigung von H9, welcher zufolge substanzielle Korrelationen zwischen den Subtestergebnissen und dem Gesamtergebnis von LESEN 6-7 und LESEN 8-9 mit der Angabe der Schüler zur Anzahl der bei ihnen zu Hause vorhandenen Bücher bestehen sollten, können ebenfalls Vorbefunde repliziert werden. Die Korrelationen fallen für den Subtest TV und das Gesamtergebnis jeweils höher aus als für den Subtest BLK und liegen für den Subtest TV und das Gesamtergebnis beider Tests im mittelhohen Bereich, während sie für den Subtest BLK beider Tests im niedrigen Bereich liegen. Mit Werten im Bereich von $r = .36$ bis $r = .50$ ist der Zusammenhang zwischen den Lesetestergebnissen und der Anzahl zu Hause vorhandener Bücher so hoch bzw. höher als der für die LGVT-Skalen berichtete Zusammenhang ($r = .41$ bzw. $r = .25$; Schneider et al., 2007).

Laut H10 sollte zudem eine substanzielle Korrelation zwischen der Häufigkeit, mit der Schüler zum Vergnügen lesen, und ihren Ergebniswerten in beiden Subtests sowie in Bezug auf das Gesamtergebnis von LESEN 6-7 und LESEN 8-9 bestehen. Dies wurde für beide Tests bestätigt: alle Korrelationen fallen signifikant aus. Die Korrelationen sind bei LESEN 6-7 für beide Subtests und das Gesamtergebnis sowie für den Subtest BLK von LESEN 8-9 niedrig, während sie für den Subtest TV und das Gesamtergebnis von LESEN 8-9 in den Bereich mittlerer Höhe fallen.

Insgesamt gehen die Korrelationshöhen der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit den Schulnoten und dem Skalenwert des Lehrerurteils größtenteils konform mit Befunden anderer Lesetestvalidierungen sowie mit Studienergebnissen zur diagnostischen Kompetenz von Lehrkräften und zu Zusammenhängen von Aspekten der Lesekompetenz mit Schulleistungen in verschiedenen Bereichen. Auch Vorbefunde zu Zusammenhängen von Lesekompetenz mit der Anzahl zu Hause vorhandener Bücher sowie mit der Häufigkeit, mit der Schüler zum Vergnügen lesen, ließen sich mit LESEN 6-7 und LESEN 8-9 replizieren. Somit spricht Vieles für die Kriteriumsvalidität der Tests.

In den vergangenen Abschnitten wurden zahlreiche Aspekte der Inhaltsvalidität, der Konstruktvalidität und der Kriteriumsvalidität betrachtet. Wenngleich die Ergebnisse teilweise nur deskriptiv beurteilt und einzelne Hypothesen nur tendenziell bestätigt werden konnten sowie einige methodische Mängel bei der Ergebnisinterpretation zu berücksichtigen sind, spricht der Großteil der Ergebnisse sowohl bei LESEN 6-7 als auch bei LESEN 8-9 für eine hohe Validität. Für erwartungswidrige Ergebnisse konnten plausible Erklärungen gefunden werden, die darauf hindeuten, dass die unerwarteten Ergebnisse nicht zwangsläufig die Validität der Tests in Frage stellen.

Kapitel 18

Zusammenfassung von Teil II

Der zweite Teil der vorliegenden Arbeit befasste sich mit der Konstruktion und Evaluation der Lesetests LESEN 6-7 und LESEN 8-9. Es wurde zunächst die Testkonstruktion von den ersten Testentwürfen über deren empirische Erprobung und mehrere Revisionen bis zu den Endversionen beschrieben. Anschließend wurden die Normierung der Endversionen, die Prüfung der Itemgüte für jeweils beide Subtests sowie die Prüfung der Rasch-Modell-Konformität für die Subtests TV dargestellt. Zuletzt erfolgte eine Überprüfung der Tests im Hinblick auf Reliabilität und Validität. Am Ende jedes Kapitels stand eine Diskussion der jeweiligen Ergebnisse.

LESEN 6-7 und LESEN 8-9 basieren inhaltlich auf den im ersten Teil der Arbeit beschriebenen theoretischen Vorstellungen und empirischen Befunden zu verschiedenen Komponenten und Ebenen des Leseverständnisses sowie zur Dimensionalität des Konstruktes Leseverständnis. Entsprechend bestehen die analog aufgebauten Tests jeweils aus zwei Subtests: dem Subtest BLK zur Erfassung der basalen Lesekompetenz und dem Subtest TV zur Erfassung des Textverständnisses. Das methodische Vorgehen bei der Testkonstruktion und der empirischen Erprobung der Endversionen von LESEN 6-7 und LESEN 8-9 erfolgte gemäß KTT sowie für den Subtest TV beider Tests zusätzlich gemäß IRT.

Der Subtest BLK ist für beide Tests identisch und besteht aus einer Satzleseaufgabe. Dieser Aufgabentyp hat sich bereits zuvor für die Erfassung der basalen Lesekompetenz in der Sekundarstufe bewährt (vgl. W. Lenhard, 2013; SLS 5-8 s. Kap. 5.3.4). Es müssen aus einer Liste von Sätzen innerhalb von drei Minuten möglichst viele Sätze still gelesen und auf inhaltliche Richtigkeit hin beurteilt werden. Dieser Subtest wurde zunächst computergestützt pilotiert, um für die Itemselektion auch die Antwortzeit für jedes einzelne Item berücksichtigen zu können. Für jedes Item wurde ein LPQ-Wert berechnet, der aus dem Quotienten „Korrektheit der Lösung durch Reaktionsgeschwindigkeit“ resultierte. Anschließend wurden auf Basis der LPQ-Werte diejenigen Items ausgewählt, die von den meisten Schülern in der kürzesten Zeit korrekt beantwortet wurden. Die Endversion des Subtests BLK enthält 100 kurze und leicht hinsichtlich inhaltlicher Richtigkeit beurteilbare Sätze. Da bei diesem Subtest die Anzahl der in der vorgegebenen Zeit korrekt beurteilten Sätze bereits eine metrische Variable darstellt, war es nicht notwendig, ein Testmodell zugrunde zu legen (vgl. J. Rost, 2004, S. 44).

Der Subtest TV besteht in beiden Tests aus jeweils zwei Texten – einem expositorischen Text und einem narrativen Text – mit einigen Verständnisfragen im SC-Format

zu jedem Text. Die Texte und Fragen unterscheiden sich bei den Tests. Bei LESEN 6-7 sind die Texte kürzer, und es werden weniger sowie leichtere Fragen gestellt als bei LESEN 8-9. Alle Items (Fragen) wurden stringent aus der Theorie abgeleitet, weshalb verschiedene Verständnisebenen sowie kontextualisiertes Wortverständnis berücksichtigt werden. Zur Itemselektion für den Subtest TV beider Tests wurden zum einen Kennwerte der KTT (Trennschärfe, Schwierigkeit, Standardabweichung und Selektionskennwert) herangezogen. Zum anderen wurde das dichotome Rasch-Modell zugrunde gelegt und es wurden IRT-Kennwerte des Item-Fits (wMNSQ) bei der Itemselektion einbezogen und darüber hinaus anhand weiterer Aspekte (Eindimensionalität, lokale Unabhängigkeit, Ratekontrolle, Personenhomogenität) die Rasch-Modell-Konformität überprüft. Zudem wurde bei der Itemselektion stets die inhaltliche Bedeutsamkeit der einzelnen Items für das Konstrukt Textverständnis beachtet. Die KTT-Kennwerte fielen in der letzten Voruntersuchung zufriedenstellend aus, während die Rasch-Kennwerte teilweise für und teilweise gegen eine Anwendbarkeit des Rasch-Modells sprachen.

Nach ihrer Fertigstellung wurden die Tests in mehreren deutschen Bundesländern normiert. Da Befunde zur Lesekompetenz deutscher Sekundarschüler zeigten, dass sich innerhalb Deutschlands verschiedene Subgruppen (z. B. Schüler verschiedener Schularten und Schüler verschiedener Bundesländer) bedeutsam in ihrer Lesekompetenz unterscheiden, wurde versucht, die Verteilung der Schüler über die Subgruppen bei der Ziehung der Normstichprobe zu berücksichtigen. Insgesamt umfassen die Normstichproben 1 644 Schüler für LESEN 6-7 und 945 Schüler für LESEN 8-9. Die Normstichproben erwiesen sich auf Klassenstufenebene als ausreichend groß, um bei Erfüllung weiterer Gütekriterien eine angemessene Differenzierung im gesamten Leistungsspektrum zu ermöglichen. Wird aber zusätzlich zur Aufteilung nach Klassenstufe auch noch nach Schulart aufgeteilt, reicht die Größe einzelner Substichproben hingegen nicht mehr aus, um in den Randbereichen gut differenzieren zu können. Im Hinblick auf die Repräsentativität der Stichproben ergab sich, dass diese trotz großen Aufwands und des Einbezugs verschiedener Schularten aus mehreren Bundesländern zwar in einigen, jedoch nicht in allen relevanten Aspekten als repräsentativ bezeichnet werden können. Für den Anteil an Schülern mit Migrationshintergrund und den Anteil an Schülern mit einer LRS-Diagnose scheint dies zumindest annähernd der Fall zu sein, während die Verteilung über die Bundesländer und Schularten nicht als repräsentativ bezeichnet werden kann. Die Frage, ob es prinzipiell möglich ist, eine repräsentative Normstichprobe zu erreichen, wurde im entsprechenden Abschnitt kritisch diskutiert.

Die Analyse der Endversionen von LESEN 6-7 und LESEN 8-9 auf Basis der Normdaten zeigte für den Subtest BLK, dass zwar in den höheren Klassenstufen die Rohwertverteilungen leicht linksschief ausfallen, aber kein ausgeprägter Deckeneffekt vorhanden ist. Die Bearbeitungszeit wurde also derart knapp gewählt, dass kaum ein Schüler alle Items korrekt bearbeiten kann. Die Betrachtung der Rohwertverteilungen der Normdaten für den Subtest TV beider Tests ergab, dass auch hier weder ein Boden-

noch ein ausgeprägter Deckeneffekt vorliegt. Die Items decken in beiden Tests jeweils ein breites Schwierigkeitsspektrum ab. Auch die übrigen auf Basis der Normdaten für die Subtests TV bestimmten KTT-Itemkennwerte fallen zufriedenstellend aus. Einzelne Selektionskennwerte sind zwar relativ niedrig ($.20 \leq SK < .30$), jedoch wurden die entsprechenden Items aufgrund ihrer inhaltlichen Bedeutung beibehalten. Zur Prüfung der Rasch-Modell-Konformität wurde die Dimensionalität anhand von Screetests, Parallelanalysen und MAP-Tests überprüft, der Item-Fit anhand von wMNSQ-Werten, die lokale Unabhängigkeit anhand des Q_3-Indexes und Personenhomogenität anhand grafischer Modelltests sowie Andersen-Tests. Mit Ausnahme des umstrittenen Andersen-Tests spricht Vieles für Rasch-Modell-Konformität beider Tests. Erklärungsmöglichkeiten für die erwartungswidrigen Ergebnisse der Andersen-Tests wurden ausführlich diskutiert.

Die Ergebnisse der Reliabilitätsanalysen fallen zufriedenstellend bis sehr gut aus. Für die Subtests TV wurden auch hier über die KTT-Methoden (Konsistenzmaße und Retestmethode) hinaus IRT-Methoden (Erwartungswertmethode und EAP/PV) herangezogen. Alle Konsistenzmaße (KR-20 und Split-Half-Maße) sprechen für eine mittelhohe bis hohe Reliabilität beider Subtests von LESEN 6-7 und LESEN 8-9. Die IRT-Kennwerte für den Subtest TV liegen auf sehr ähnlicher Höhe. Die Retestkennwerte bei einem Wiederholungsintervall von drei Wochen fallen alle in den Bereich mittlerer Höhe, wobei zum Teil nicht unbedeutende Wiederholungseffekte auftraten, die diskutiert wurden. Insgesamt scheint die Reliabilität sowohl für eine Beurteilung von Gruppendifferenzen und interindividuellen Differenzen als auch für eine Individualdiagnostik ausreichend zu sein. Lediglich bei einer Aufteilung nach Klassenstufe und Schulart fallen einzelne Werte niedrig aus, was vermutlich der dann sehr kleinen Stichproben geschuldet ist.

Ebenso zufriedenstellend fallen die Ergebnisse der Validitätsanalysen aus. Die inhaltliche Validität wurde hauptsächlich aufgrund theoretischer Überlegungen geprüft. Beim Subtest BLK wurden zusätzlich die Items auf Personen- und Itemebene betrachtet, wobei die Ergebnisse darauf hindeuten, dass zum einen die Schüler die Items instruktionsgemäß bearbeiteten, und dass zum anderen die Items leicht genug sind und die Bearbeitungsdauer knapp genug ist, um aufgrund der basalen Lesekompetenz und nicht aufgrund höherer Verständnisleistungen oder aufgrund von Vorwissen zu differenzieren. Die inhaltliche Validität der Subtests TV wurde mit der streng theoriegeleiteten Itemgenerierung begründet und damit, dass die inhaltliche Bedeutung der Items für das Konstrukt Leseverständnis bei der Itemselektion stets beachtet wurde.

Zur Prüfung der Konstruktvalidität wurden die Ergebnisse von LESEN 6-7 und LESEN 8-9 mit Ergebnissen verschiedener konstruktnaher Skalen (LGVT-LG, LGVT-LV, WLST 7-12) korreliert. Von den 27 berechneten Korrelationen fallen 26 Korrelationen signifikant aus. Korrelationen mit Ergebnissen von Skalen, die genau das gleiche Konstrukt erfassen wie der entsprechende Subtest von LESEN 6-7 und LESEN 8-9, fallen hoch bis sehr hoch aus. Weiter wurden die Ergebnisse von LESEN 6-7 und LESEN 8-9 zur

Prüfung der Kriteriumsvalidität mit den Skalenwerten verschiedener Außenkriterien korreliert. Dabei wurden neben konstruktnahen Kriterien (Lehrerurteil, Deutschnote) auch konstruktferne Kriterien (Gesamtnotenschnitt, Mathematiknote) berücksichtigt, um nicht nur konvergente, sondern auch diskriminante Validität zu prüfen. Es zeigt sich u. a. für beide Tests für 23 der 24 Korrelationen das erwartete Muster für die Korrelationen der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit den Kriteriumswerten: Korrelation mit dem Skalenwert des Lehrerurteils $>$ Korrelation mit der Deutschnote $>$ Korrelation mit dem Gesamtnotenschnitt $>$ Korrelation mit der Mathematiknote.

Sowohl im Rahmen der Prüfung der Konstruktvalidität als auch im Rahmen der Prüfung der Kriteriumsvalidität wurden darüber hinaus Hypothesen geprüft, die auf empirischen Vorbefunden und theoretischen Überlegungen basierten. Die Ergebnisse in Bezug auf Klassenstufen-, Schulart- und Geschlechterunterschiede fallen ebenso größtenteils erwartungskonform aus wie die Ergebnisse in Bezug auf Unterschiede zwischen Schülern mit LRS-Diagnose und Schülern ohne LRS-Diagnose sowie die Ergebnisse in Bezug auf Unterschiede zwischen Schülern mit Deutsch als Muttersprache und Schülern mit einer anderen Muttersprache als Deutsch. Des Weiteren erwiesen sich der Zusammenhang von Leseverständnis mit der Häufigkeit, mit der Schüler zum Vergnügen lesen, sowie der Zusammenhang von Leseverständnis mit der Anzahl der zu Hause vorhandenen Bücher erwartungskonform als signifikant. Diese somit bis auf einzelne Ausnahmen erwartungskonformen Ergebnisse sind ein weiterer Hinweis auf die Validität von LESEN 6-7 und LESEN 8-9. Insgesamt scheinen sowohl Inhalts- als auch Konstrukt- und Kriteriumsvalidität für beide Tests gegeben zu sein. Einzelne hypothesenwidrige Ergebnisse wurden in den entsprechenden Abschnitten ausführlich diskutiert und stets plausible Erklärungsmöglichkeiten dargelegt.

Abbildung 23 fasst alle im Rahmen der Testkonstruktion und empirischen Erprobung verwendeten Kennwerte und durchgeführten Tests in einer Übersicht zusammen. Schlussendlich fallen die Ergebnisse der empirischen Erprobung von LESEN 6-7 und LESEN 8-9 sehr zufriedenstellend aus.

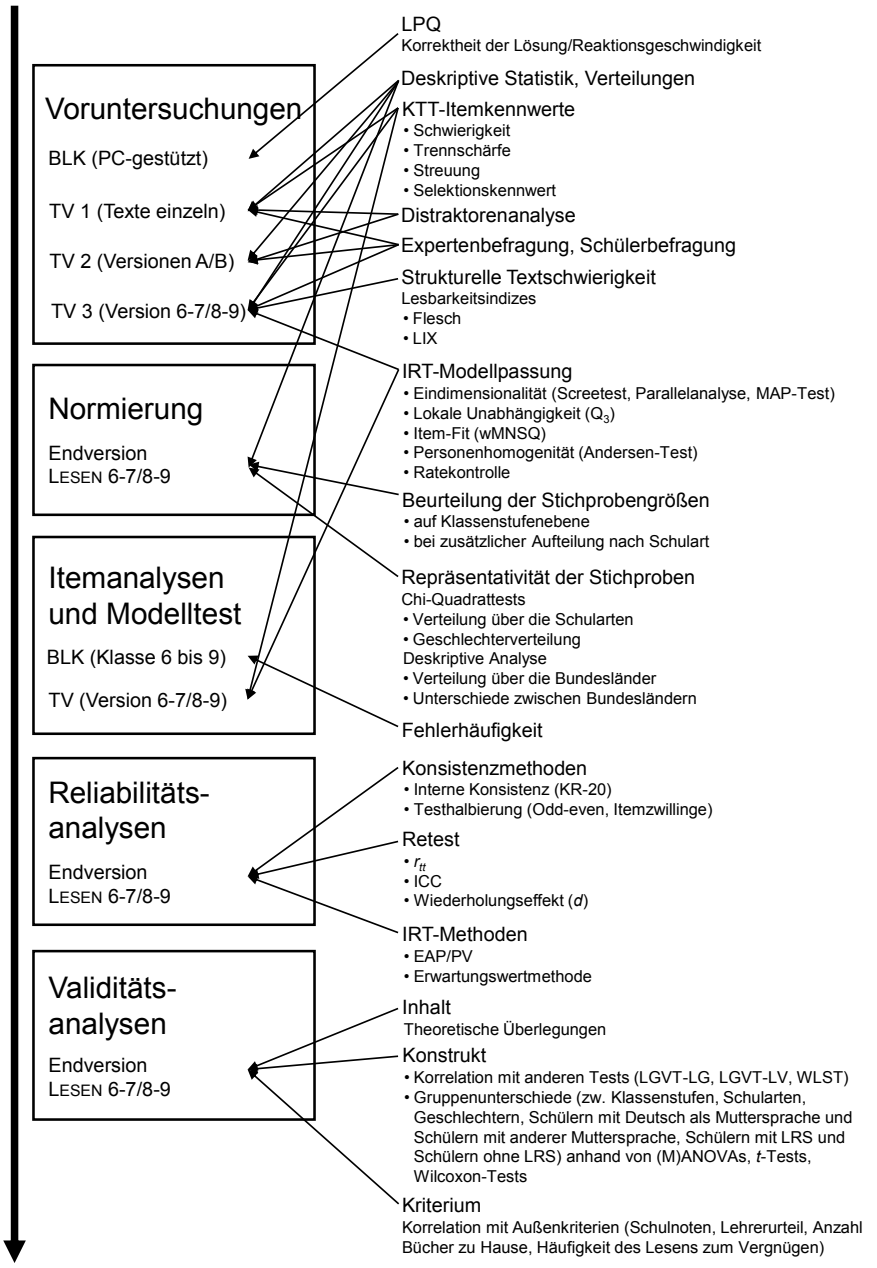


Abbildung 23. Übersicht aller im Rahmen der Testkonstruktion und empirischen Erprobung verwendeten Kennwerte und durchgeführten Tests.

Teil III

Abschließende Diskussion

Der erste Teil der vorliegenden Arbeit beschrieb die theoretischen Grundlagen und empirischen Befunde, auf welchen die im zweiten Teil beschriebene Konstruktion und empirische Erprobung von LESEN 6-7 und LESEN 8-9 basierte. Der abschließende dritte Teil besteht nun aus einer Bewertung der Endversionen von LESEN 6-7 und LESEN 8-9 anhand der im ersten Teil der Arbeit in Kapitel 5.3.2 beschriebenen allgemein anerkannten Testgütekriterien. Zudem werden einige Kritikpunkte am methodischen Vorgehen und an den Endversionen besprochen, und es wird auf sich aus den Kritikpunkten ergebende Ansatzpunkte und Fragestellungen für zukünftige Forschungsbemühungen eingegangen. Das Fazit fasst die Stärken und Schwächen von LESEN 6-7 und LESEN 8-9 zusammen.

Kapitel 19

Bewertung der Endversionen anhand der Testgütekriterien

Wie im ersten Teil der Arbeit (Kap. 5.3.2) dargestellt, existieren zahlreiche allgemein anerkannte Testgütekriterien, die ein psychologischer Test erfüllen sollte. Dazu zählen neben den Hauptgütekriterien Objektivität, Reliabilität, Validität und Skalierbarkeit auch die Nebengütekriterien Normierung, Vergleichbarkeit, Ökonomie, Nützlichkeit, Zumutbarkeit, Fairness und Nicht-Verfälschbarkeit. Im Folgenden werden LESEN 6-7 und LESEN 8-9 im Hinblick auf alle diese Kriterien bewertet.

Objektivität. Sofern alle Anweisungen des Testmanuals befolgt werden, können LESEN 6-7 und LESEN 8-9 hinsichtlich Durchführung, Auswertung und Interpretation als objektiv bezeichnet werden. *Durchführungsobjektivität* wird erreicht, indem sich der Testleiter genau an die Anweisungen im Testmanual hält und die Instruktionen wörtlich wiedergibt. Das Testmanual enthält zudem genaue Vorgaben in Bezug auf die Durchführungsbedingungen, die Zeitbegrenzungen für die Subtests sowie Hinweise zur Beantwortung von Fragen der Schüler während der Testung. *Auswertungsobjektivität* ist durch explizite und präzise Auswertungsvorschriften sichergestellt. Die korrekten Antworten sind eindeutig festgelegt und aufgrund des gebundenen Antwortformats und der verfügbaren Auswertungsschablonen leicht zu ermitteln. Trotz der Intuitivität der Auswertungshilfen ist das Vorgehen zur Ermittlung der Roh- und Normwerte im Testmanual genau erläutert und die Erstellung eines Profildiagramms anhand eines Beispiels veranschaulicht. *Interpretationsobjektivität* ergibt sich durch einen Bezug auf die Normstichprobe, also aus dem Vergleich der Ergebnisse der Testpersonen mit den Ergebnissen der Normstichprobe. Hierfür stehen Normtabellen zur Verfügung, deren Nutzung im Testmanual beschrieben ist. Zudem enthält das Testmanual ausführliche Erläuterungen zur Interpretation der Roh- und Normwerte, die sowohl grafisch veranschaulicht als auch verbalisiert dargestellt sind.

Reliabilität. Zur Prüfung der Reliabilität von LESEN 6-7 und LESEN 8-9 wurden sowohl KTT-Kennwerte als auch – allerdings nur für den Subtest TV – IRT-Kennwerte herangezogen. Konsistenzwerte für beide Subtests sowie auch die Ergebnisse der IRT-Methoden für den Subtest TV sprechen für eine hohe Reliabilität beider Subtests von LESEN 6-7 und LESEN 8-9. Die Werte der Retestkorrelationen bei einem Wiederholungsintervall von drei Wochen, die sowohl für beide Subtests als auch für das Gesamt-

ergebnis ermittelt wurden, fallen in einen Bereich mittlerer Höhe. Insgesamt sprechen die Ergebnisse dafür, dass die Reliabilität sowohl für die Beurteilung von Gruppendifferenzen und interindividuellen Differenzen als auch für die Individualdiagnostik ausreichend ist.

Validität. Auf inhaltliche Validität von LESEN 6-7 und LESEN 8-9 kann aufgrund theoretischer und sachlich-logischer Überlegungen geschlossen werden. Beim Subtest BLK zeigt sich bei einer Betrachtung der Normdaten auf Personen- und Itemebene, dass zum einen die Schüler den Subtest instruktionsgemäß bearbeiteten, und dass zum anderen die Items leicht sind und die Bearbeitungsdauer angemessen ist, um aufgrund der basalen Lesekompetenz und nicht aufgrund höherer Verständnisleistungen oder aufgrund von Vorwissen zu differenzieren. Die inhaltliche Validität des Subtests TV beider Tests kann mit den stringent aus der Theorie abgeleiteten Items begründet werden sowie damit, dass die inhaltliche Bedeutung der Items für eine umfassende Abbildung des Konstrukts Leseverständnis bei der Itemselektion berücksichtigt wurde.

Für Konstruktvalidität sprechen die (bis auf eine Ausnahme) signifikanten Korrelationen der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit Ergebnissen konstruktnaher Skalen (LGVT-LG, LGVT-LV und WLST 7-12) sowie insbesondere die hohen bis sehr hohen Korrelationen mit Ergebnissen von Skalen, die genau das gleiche Konstrukt erfassen wie der entsprechende Subtest von LESEN 6-7 und LESEN 8-9. Auch die Replikation zahlreicher – wenn auch nicht aller – Vorbefunde bezüglich Unterschieden zwischen verschiedenen Subgruppen der Zielpopulation (Unterschiede zwischen den Klassenstufen, den Schularten und den Geschlechtern, Unterschiede zwischen Schülern mit und ohne Migrationshintergrund sowie zwischen Schülern mit und ohne LRS-Diagnose) weisen auf Konstruktvalidität von LESEN 6-7 und LESEN 8-9 hin. Für die nicht erwartungskonformen Ergebnisse ließen sich plausible Erklärungen finden, die dafür sprechen, dass die der Erwartung widersprechenden Ergebnisse nicht zwingend auf mangelnde Validität von LESEN 6-7 bzw. LESEN 8-9 zurückzuführen sein müssen.

Weiter fallen die Korrelationen der Ergebnisse von LESEN 6-7 und LESEN 8-9 mit konstruktnahen Kriteriumswerten (Lehrerurteil, Deutschnote) höher aus als die Korrelationen mit konstruktfernen Kriteriumswerten (Gesamtnotenschnitt, Mathematiknote). Für beide Tests zeigte sich bei diesen Korrelationen größtenteils das erwartete Muster: Korrelation mit dem Skalenwert des Lehrerurteils $>$ Korrelation mit der Deutschnote $>$ Korrelation mit dem Gesamtnotenschnitt $>$ Korrelation mit der Mathematiknote. Diese Ergebnisse sprechen sowohl für konvergente Validität als auch für diskriminante Validität. Darüber hinaus konnten mit LESEN 6-7 und LESEN 8-9 empirische Vorbefunde in Bezug auf den Zusammenhang von Leseverständnis mit der Häufigkeit, mit der Schüler zum Vergnügen lesen, und der Anzahl der Bücher, die bei den Schülern zu Hause verfügbar ist, repliziert werden, was ebenfalls für die Validität der Tests spricht.

Skalierbarkeit. Beim Subtest BLK stellt die Anzahl der in der vorgegebenen Zeit gelesenen und korrekt beurteilten Sätze eine metrische Variable dar. Somit weisen die Ergebnisse Intervallskalenniveau auf. Zudem sprechen die geringen Fehlerzahlen der einzelnen Versuchspersonen der Normstichprobe dafür, dass die Schüler den Subtest instruktionsgemäß bearbeiteten und reines Raten kaum eine Rolle spielte. Bei sehr vielen Fehlern oder ausgelassenen Items bei einer Testperson, ist jedoch die Aussagekraft des Testergebnisses anzuzweifeln, da möglicherweise durch Raten eine unverhältnismäßig hohe Punktzahl erreicht wurde. Bei instruktionsgemäßer Bearbeitung stellt die Summenbildung aus den Itemwerten für den Subtest BLK jedoch eine gültige Verrechnungsvorschrift dar.

Für den Subtest TV beider Tests ist das Kriterium der Skalierbarkeit durch die Zugrundelegung des dichotomen Rasch-Modells erfüllt. Mit Ausnahme der umstrittenen und kritisch diskutierten Andersen-Tests sprechen die Ergebnisse der empirischen Erprobung der Subtests TV dafür, dass die Items die Annahmen des dichotomen Rasch-Modells (Eindimensionalität, Gleichheit der Trennschärfen, kein bedeutsamer Einfluss von Raten, lokale Unabhängigkeit) erfüllen. Somit kann für die Ergebnisse Intervallskalenniveau angenommen werden, das Aufsummieren der Itemwerte zu einem Ergebniswert für den Subtest TV stellt eine gültige Verrechnungsvorschrift dar, und der Summenwert ist eine suffiziente Statistik.

Normierung. Beide Tests wurden jeweils anhand einer Stichprobe normiert, welche Schüler aus Hauptschulen, Realschulen und Gymnasien sieben deutscher Bundesländer einschließt. Insgesamt umfassen die Normstichproben 1 644 Schüler für LESEN 6-7 und 945 Schüler für LESEN 8-9. Die Normstichproben sind auf Klassenstufenebene ausreichend groß, um bei Erfüllung weiterer Gütekriterien im gesamten Leistungsspektrum der Zielgruppen differenzieren zu können. Wird aber zusätzlich noch nach Schulart aufgeteilt, reicht die Größe der Substichproben nicht in allen Fällen aus, um auch in den Randbereichen angemessen differenzieren zu können. Im Hinblick auf die Repräsentativität der Stichproben zeigte sich, dass diese trotz großen Aufwands und trotz des Einbezugs verschiedener Schularten aus mehreren Bundesländern zwar bezüglich einiger, jedoch nicht aller Aspekte, die sich als relevant erwiesen haben, als repräsentativ gelten können.

Im Anhang des Testmanuals werden Tabellen mit Normwerten zur Verfügung gestellt, in denen jedem erreichbaren Rohwert ein T-Wert, ein T-Wertband sowie ein Prozentrang zugeordnet sind. Für den Subtest TV gibt es darüber hinaus für jeden Test eine Tabelle, in der jedem erreichbaren Rohwert ein Personenparameter zugeordnet ist. Das Kriterium der Normierung ist für LESEN 6-7 und LESEN 8-9 somit erfüllt.

Vergleichbarkeit. Da keine parallelen Testformen vorliegen (vgl. 5.3.4), wird das Kriterium der Vergleichbarkeit von LESEN 6-7 und LESEN 8-9 nicht erfüllt. Auf eine Erstellung paralleler Tests wurde aufgrund des unverhältnismäßig hohen Aufwands verzichtet.

Ökonomie. Die Durchführung der Tests dauert jeweils eine ganze Schulstunde und somit länger als die Durchführung einiger anderer Lesetests für die Sekundarstufe (z. B. LGVT 6-12 oder SLS 5-8, vgl. 5.3.4). Betrachtet man jedoch den Nutzen – man erhält ein Leseprofil, das basale Lesekompetenzen und Textverständnis beinhaltet und beim Textverständnis sogar höhere Verständnisebenen einschließt – so erscheint der Aufwand im Verhältnis dazu angemessen. Zudem ist der Test als Gruppentest einsetzbar, was einen Vorteil gegenüber anderen Lesetests darstellt, die das Leseverständnis erfassen (z. B. ZLT-II, vgl. 5.3.4). Mithilfe von Schablonen sind LESEN 6-7 und LESEN 8-9 darüber hinaus schnell und einfach auswertbar. Dank des gebundenen Antwortformats ist sogar eine Auswertung anhand eines Dokumentenscanners möglich, was sich vor allem bei einer Auswertung einer großen Anzahl an Testheften z. B. im Rahmen von Forschungsprojekten vorteilhaft nutzen lässt. Insgesamt lassen sich die Tests also sowohl hinsichtlich der Durchführung als auch hinsichtlich der Auswertung als ökonomisch bezeichnen.

Nützlichkeit. LESEN 6-7 und LESEN 8-9 können als nützlich bezeichnet werden, da sie sich in einigen Punkten von bereits verfügbaren Lesetests abheben: Im Gegensatz zu bereits verfügbaren Tests für die Zielgruppe – z. B. Ein-Minuten-Leseflüssigkeitstest, SLS 5-8, LGVT 6-12 – stellen LESEN 6-7 und LESEN 8-9 nicht nur Screenings dar und erfassen nicht nur basale Lesekompetenzen und textbasiertes Leseverständnis, sondern zusätzlich auch noch ein tiefergehendes Textverständnis einschließlich der Generierung von Inferenzen und des Aufbaus eines Situationsmodells. LESEN 6-7 und LESEN 8-9 können dadurch in den angegebenen Klassenstufen jeweils im gesamten Leistungsspektrum angemessen differenzieren. Hierfür war bislang kein Test verfügbar. Ein weiterer Vorteil (z. B. gegenüber dem ebenfalls recht umfassenden ZLT-II) ist die Möglichkeit, LESEN 6-7 und LESEN 8-9 als Gruppentest einzusetzen. Darüber hinaus wurden LESEN 6-7 und LESEN 8-9 jeweils an einer großen Stichprobe normiert, die mehrere Bundesländer umfasst. Dies ist ein Vorteil gegenüber vielen verfügbaren Tests, z. B. FLVT 5-6 und LEVE (vgl. Kap. 5.3.4). LEVE kann zudem nur computergestützt durchgeführt werden, was den Einsatz in Schulklassen, die zu einem großen Teil immer noch spärlich mit PCs ausgestattet sind, schwierig macht. LESEN 6-7 und LESEN 8-9 liegen dagegen in Papierform vor und sind somit in jeder Schule im normalen Klassenzimmer einsetzbar.

Im Gespräch mit Lehrkräften im Rahmen der Voruntersuchungen und der Normierung von LESEN 6-7 und LESEN 8-9 wurde von den Lehrkräften immer wieder betont, dass sie die Tests für sinnvoll und hilfreich halten, da es ihnen im Alltag oft ein Rätsel sei, wie die schlechten Leseverständnisleistungen ihrer Schüler zustande kommen. Auch die Befunde zur zum Teil mangelhaften diagnostischen Kompetenz von Lehrkräften in der Sekundarstufe im Hinblick auf die Lesekompetenz ihrer Schüler (vgl. Kap. 5.2) sprechen dafür, dass Lehrkräfte von derartigen Tests profitieren könnten. LESEN 6-7 und LESEN 8-9 liefern neben einem Gesamtergebnis auch ein Ergebnis, das

über die basale Lesekompetenz informiert, und zusätzlich noch ein Ergebnis zur Textverständnisleistung. Aus der Ermittlung von Stärken und Schwächen eines Schülers ergeben sich dabei Anhaltspunkte für die Auswahl adäquater Fördermaßnahmen. Der Subtest TV von LESEN 6-7 und LESEN 8-9 enthält zudem einen expositorischen und einen narrativen Text, womit sich die Möglichkeit ergibt, es zu erkennen, falls bei einem Schüler besonders schwache Leistungen bei der Bearbeitung einer der beiden Textgenres vorliegen. In diesem Fall kann ebenfalls auf entsprechende Maßnahmen zur Unterstützung der Sachtextlektüre bzw. zur Unterstützung des literarischen Lesens zurückgegriffen werden (vgl. Kap. 8). Extrem schwache Leistungen bei LESEN 6-7 oder LESEN 8-9 können auf schwerwiegendere Leseprobleme hindeuten. Vor allem stark unterdurchschnittliche Ergebnisse in beiden Subtests und ein damit einhergehendes besonders schwaches Gesamtergebnis können ein Hinweis für das Vorliegen einer LRS sein (vgl. Kap. 17). Die Testmanuale stellen für jedes der genannten Defizite zusätzlich Informationen bereit und machen Empfehlungen, welche Fördermaßnahmen bei welchen Defiziten sinnvoll sind. Artelt et al. (2007) fordern in einer BMBF-Expertise zur Leseförderung zusätzlich zu Motivierungsmaßnahmen stärker auf einzelne Teilkomponenten des Leseverständnisses und die Prozesse der Bedeutungskonstruktion ausgerichtete Fördermaßnahmen einzusetzen. Die Ergebnisse von LESEN 6-7 und LESEN 8-9 können hierfür eine Grundlage bieten.

Zumutbarkeit. Mit einer Dauer von einer Schulstunde ist der Aufwand für Schüler der Regelschulen zumutbar. Die Bearbeitungsdauer ist streng begrenzt, und es handelt sich um eine Zeitspanne, an die die Schüler aus dem Schulalltag gewöhnt sind und die sie überschauen können. Es sind sowohl sehr leichte als auch sehr schwere Items enthalten, sodass jeder Schüler zumindest einen Teil der Aufgaben lösen können sollte. In den Voruntersuchungen wurden die Schüler unter anderem auch zur Zumutbarkeit von LESEN 6-7 und LESEN 8-9 befragt. Die meisten Schüler bewerteten die Anforderungen als angemessen und zumutbar.

Fairness. Die Ergebniswerte beider Subtests sind metrisch skaliert, sodass unterschiedliche Testergebnisse tatsächliche Unterschiede in der Leistung widerspiegeln, was eine wichtige Grundlage für eine faire Messung darstellt. Um eine faire Ergebnisinterpretation zu erleichtern, werden zudem klassenstufen- und schulartspezifische Vergleichswerte zur Verfügung gestellt. Zudem wurde versucht, eine möglichst repräsentative Normstichprobe zu erzielen, die beispielsweise Schüler aus verschiedenen Bundesländern einschließt. Weiter wurde bei der Konstruktion des Subtests TV darauf geachtet, Texte auszuwählen, die Themen behandeln, welche nicht typischerweise von einem Geschlecht bevorzugt werden. Bei der Auswertung der Normdaten ergaben sich nur in Teilbereichen signifikante Geschlechterunterschiede, und diese wiesen nur eine niedrige Effektstärke auf, sodass sie nicht praktisch bedeutsam sind.

Schüler mit LRS – und im Subtest TV auch Schüler mit einer anderen Muttersprache als Deutsch – zeigen zwar schlechtere Leistungen als Schüler ohne LRS bzw. Schüler mit Deutsch als Muttersprache, jedoch wurden hier keine getrennten Normen erstellt, da die Schwäche mit dem Test erkannt und anschließend gezielte Fördermaßnahmen ergriffen werden sollen. Dies entspricht den Bedingungen in der Schule, in der Arbeitswelt und im Alltag, wo ebenfalls gleiche Anforderungen unabhängig von Geschlecht, Muttersprache und LRS erfüllt werden müssen.

Nicht-Verfälschbarkeit. Bei Leistungstests lässt sich das Ergebnis generell nur nach unten verfälschen (J. Rost, 2004, S. 43). Im Subtest BLK soll die Bewertung von falsch angekreuzten und ausgelassenen Items mit null Punkten verhindern, dass Schüler die Sätze zugunsten der Geschwindigkeit gar nicht oder nur sehr ungenau lesen. Sollte jedoch von einem Schüler bei sehr vielen Items die falsche Antwortalternative angekreuzt worden sein, kann dies ein Hinweis darauf sein, dass ein Schüler die Sätze sehr ungenau gelesen und möglicherweise sogar geraten hat. Die Wahrscheinlichkeit, durch reines Raten auf die richtige Lösung zu kommen, liegt bei jedem Item bei 50 %. Dadurch kann ein Schüler, der bei allen Items rät, insgesamt eine relativ hohe Punktzahl erreichen im Vergleich zu einem schwachen Schüler, der die Items instruktionsgemäß bearbeitet und nur weniger als die Hälfte der Items zu bearbeiten schafft. Dies ist bei der Ergebnisinterpretation zu berücksichtigen.

Im Subtest TV ist reines Raten durch die relativ große Anzahl von Antwortoptionen wenig erfolgsversprechend. Weiter kann bei einer sachgerechten Durchführung unter standardisierten Bedingungen kein Abschreiben von anderen Schülern stattfinden. Auch die Konformität der Items mit dem dichotomen Rasch-Modell spricht dafür, dass reines Raten eine zu vernachlässigende Rolle spielt.

Insgesamt fällt die Bewertung von LESEN 6-7 und LESEN 8-9 in Bezug auf die Testgütekriterien überwiegend positiv aus. Die Tests erfüllen fast alle Kriterien. Ein ausführliches Manual und Auswertungsschablonen führen dazu, dass die Objektivität der Tests bei Einhaltung der im Manual beschriebenen Vorgaben gewährleistet ist. Die Kriterien der Reliabilität und Validität wurden im Verlauf dieser Arbeit ausführlich geprüft, und die Ergebnisse sprechen größtenteils für eine sehr zufriedenstellende Erfüllung beider Kriterien. Die meisten Ergebnisse der Prüfung auf Rasch-Modell-Konformität bestätigen diese, weshalb Skalierbarkeit angenommen werden kann. Weiter sind die Tests ökonomisch, da sie im Verhältnis zum Durchführungs- und Auswertungsaufwand umfangreiche und nützliche Informationen liefern. Dass sich die Durchführungsdauer auf lediglich eine Schulstunde beschränkt, wirkt sich zudem positiv auf die Zumutbarkeit aus. Im Hinblick auf die Fairness werden beispielsweise für verschiedene Subgruppen getrennte Normtabellen zur Verfügung gestellt, und die Tests sind nur in geringem Maße verfälschbar. Das Kriterium der Vergleichbarkeit ist dagegen nicht erfüllt, da keine parallelen Tests zur Verfügung stehen. Das Kriterium der Normierung ist nur in-

sofern erfüllt, dass eine Normierung an großen und heterogenen Stichproben stattgefunden hat und entsprechende Vergleichswerte zur Verfügung gestellt werden. Jedoch können die Normstichproben selbst nicht als repräsentativ für die gesamte Zielpopulation bezeichnet werden.

Kapitel 20

Kritik und Ausblick

Wie bereits in den Diskussionen am Ende der einzelnen Kapitel in Teil II deutlich wurde, mussten bei der Testkonstruktion sowie bei der empirischen Erprobung der Endversionen von LESEN 6-7 und LESEN 8-9 zahlreiche Kompromisse zwischen dem theoretisch Wünschenswerten und dem praktisch Umsetzbaren eingegangen werden. Aus den Kritikpunkten und offen gebliebenen Fragen ergeben sich weitere Fragestellungen und Anregungen für Forschungsbemühungen. Die wichtigsten davon sollen im Folgenden kurz dargelegt werden.

Zunächst ist zur Rasch-Skalierung kritisch anzumerken, dass sowohl für LESEN 6-7 als auch für LESEN 8-9 der Andersen-Test signifikant ausfällt, was gegen ein Vorliegen von Personenhomogenität spricht. Dabei bleibt offen, wie diese Ergebnisse zu bewerten sind. Die diesbezüglichen Empfehlungen in der Literatur sind widersprüchlich. Es scheint klar zu sein, dass der Andersen-Test sehr sensitiv ist und bei angemessen großen Stichproben auch schon bei minimalen Abweichungen vom Modell signifikant ausfällt. Einige Autoren halten die Durchführung eines derartigen Tests dennoch für unerlässlich und bewerten ein signifikantes Ergebnis als eindeutigen Hinweis darauf, dass die Tests nicht Rasch-Modell-konform sind (z. B. K. D. Kubinger, 2006; Bühner, 2011; Strobl, 2010). Andere Autoren beschränken sich dagegen auf die Prüfung des Item-Fits und erwähnen globale Modelltests gar nicht erst, oder sie äußeren sich sogar kritisch dazu (z. B. DeMars, 2010, S. 57ff.). Es scheint also noch weitere Forschung nötig zu sein, um konsistente Empfehlungen für die Praxis der Testkonstruktion und Ergebnisbewertung ableiten zu können. Möglicherweise wäre die Berechnung einer Effektgröße hilfreich, die eine von der Stichprobengröße unabhängige Ergebnisbewertung erlaubt.

Hinsichtlich der Dimensionalität von LESEN 6-7 und LESEN 8-9 zeigte sich in der vorliegenden Arbeit zum einen, dass die Korrelation zwischen dem Subtest BLK und dem Subtest TV jeweils nur in den Bereich mittlerer Höhe fällt, und zum anderen sprechen die Ergebnisse trotz der Berücksichtigung verschiedener Textgenres sowie verschiedener Verständnisebenen für eine Eindimensionalität der Subtests TV. Die Ergebnisse gehen somit konform mit dem aktuellen Forschungsstand, dass Leseverständnis sich aus den beiden Komponenten „basale Lesekompetenz“ und „Leseverständnis“ zusammensetzt, wobei beide Komponenten als nicht unabhängig voneinander, aber auch nicht sehr eng zusammenhängend betrachtet werden (vgl. Schneider, 2008). Die Eindimensionalität des Subtests TV entspricht ebenfalls dem aktuellen Kenntnisstand,

dass sich Leseverständnis nicht in weitere distinkte Skalen untergliedern lässt (vgl. D. H. Rost & Buch, 2010). Jedoch ist zu berücksichtigen, dass in der vorliegenden Arbeit die anfänglich angestrebte Gleichverteilung der Items über die verschiedenen Verständnisebenen im Verlauf der Testkonstruktion zugunsten der statistischen Itemgüte aufgegeben wurde und dabei vor allem Items eliminiert wurden, die sich auf die Metaebenen des Leseverständnisses bezogen. Dies könnte ein Hinweis darauf sein, dass es sich dabei also doch um eine weitere distinkte Dimension von Leseverständnis handelt. Diese Hypothese wäre empirisch zu prüfen. Sollte sie sich bestätigen, könnte es sinnvoll sein, die Metaebene anhand einer separaten Skala zu erheben.

Laut D. H. Rost und Buch (2010) und auch im Sinne des SVR wäre es für eine adäquate Beurteilung des Leseverstehens zudem generell erforderlich, zusätzlich das Hörverstehen zu berücksichtigen. Dies war bei LESEN 6-7 und LESEN 8-9 nicht möglich, da die Tests als Gruppentests konzipiert sein sollten. Auch eine zusätzliche Erfassung des Verbal-IQs wäre im angestrebten Zeitrahmen von einer Schulstunde nicht möglich gewesen. Sollte es sinnvoll erscheinen, das rezeptive Sprachverständnis bzw. die verbale Intelligenz zusätzlich zu erfassen, muss auf andere entsprechende Verfahren zurückgegriffen werden. Grundsätzlich wäre auch die Konstruktion einer Skala zur Erfassung des Hörverstehens denkbar, die dann bei Bedarf im Rahmen einer Individualdiagnostik als optionale Ergänzung zu LESEN 6-7 und LESEN 8-9 eingesetzt werden könnte.

Schließlich kann bei LESEN 6-7 und LESEN 8-9 ähnlich wie bei den PISA-Lesetests kritisiert werden, dass sie sich auf die kognitiven Aspekte des Lesens beschränken und somit didaktisch wenig wertvolle Informationen liefern. Didaktisch und interventionsorientierten Modellen (z. B. Joshi & Aaron, 2000; Hurrelmann, 2006) zufolge müssten über die kognitiven Aspekte des Lesens hinaus auch der psychologische Bereich und Umwelteinflüsse berücksichtigt werden. Diese erweiterten Modelle lassen sich jedoch schwer operationalisieren. Außerdem wurde im Rahmen der vorliegenden Arbeit dargestellt, dass sich aus den Ergebnissen von LESEN 6-7 und LESEN 8-9 durchaus angemessene Fördermaßnahmen ableiten lassen. Hier könnten dennoch ebenfalls Verfahren entwickelt und optional eingesetzt werden, die ergänzend zu LESEN 6-7 und LESEN 8-9 den psychologischen Bereich und Umwelteinflüsse abdecken, um bei Bedarf ein umfassenderes Bild der Lesekompetenz im weiteren Sinne zu erhalten.

Aus der Tatsache, dass die Normierung, die Validierung und die weitere empirische Erprobung von LESEN 6-7 und LESEN 8-9 in deutschen Regelschulen stattfanden, ergab sich einerseits ein hohes Maß an ökologischer Validität, andererseits ergaben sich daraus auch einige Restriktionen. So gibt es in allen Bundesländern von den zuständigen Behörden bestimmte Vorgaben, ob und wie Datenerhebungen in Schulen stattfinden können. Auch die Freiwilligkeit der Teilnahme führt zu einer nicht zu vernachlässigenden Selbstselektion und beeinträchtigt die Repräsentativität der Normstichproben. Die Frage, ob es also möglich ist, eine repräsentative Normstichprobe für einen Schultest zu erhalten, wurde im entsprechenden Abschnitt der vor-

liegenden Arbeit kritisch diskutiert. Verschiedene Möglichkeiten, die Repräsentativität nachträglich zu erhöhen, wurden ebenfalls behandelt. Es zeigte sich jedoch, dass keine der Methoden wirklich zufriedenstellend ist. Die Bestimmung von Niveaustufen ist eine Möglichkeit, den Vergleich mit einer Normstichprobe zu umgehen und die Leistung von Schülern anhand inhaltlicher Kriterien zu bewerten. Auch diesbezüglich zeigen sich allerdings gravierende Nachteile, die diese Möglichkeit insbesondere für die Individualdiagnostik praktisch unbrauchbar machen (vgl. Kap. 5.3.2.2). Sowohl die Möglichkeiten der normorientierten Ergebnisinterpretation als auch die Möglichkeiten der Nutzung von Niveaustufenmodellen bzw. generell kriteriumsorientierter Ergebnisinterpretation sind daher kritisch zu sehen. Hier ist für die Zukunft noch an zufriedenstellenderen Lösungen zu arbeiten.

Weiter wäre für einen aussagekräftigen und zufallskritisch absicherbaren Vergleich der Bundesländer bezüglich der Lesekompetenz eine umfangreichere Datenerhebung erforderlich gewesen, als sie im Rahmen der vorliegenden Arbeit stattfand. Hierfür hätten in allen Bundesländern – idealerweise per Zufall – mehrere Schulen jeder Schulart ausgewählt werden müssen, damit Schulzugehörigkeit, Schulartzugehörigkeit und Bundeslandzugehörigkeit nicht konfundiert sind. Dies erscheint jedoch aufgrund der bereits angesprochenen Vorgaben der zuständigen Behörden und aufgrund der Freiwilligkeit der Teilnahme von Schulen, Lehrkräften und Schülern kaum zu bewerkstelligen. In jedem Bundesland mehrere Klassen pro Klassenstufe und Schulart zu rekrutieren, sollte jedoch prinzipiell – wenn auch mit großem Aufwand – umsetzbar sein.

Darüber hinaus sind insbesondere die Ergebnisse bezüglich der Unterschiede zwischen Schülern mit Deutsch als Muttersprache und Schülern mit einer anderen Muttersprache als Deutsch in der vorliegenden Arbeit mit äußerster Vorsicht zu interpretieren, da die Stichproben aufgrund vieler fehlender oder nicht verwertbarer Angaben sehr klein waren und vermutlich unter anderem aufgrund der Selbstselektion nicht als repräsentativ bezeichnet werden können. Auf eine Berücksichtigung des familiären Hintergrunds (SÖS und Bildungshintergrund) sollte jedoch wegen der Konfundierung mit der Muttersprache nicht verzichtet werden. Anhand eines Elternfragebogens könnten einerseits in künftiger Forschung zuverlässigere und vollständigere Daten erhoben werden, andererseits könnte dies den verzerrenden Einfluss der Selbstselektion noch erhöhen, da vermutlich viele Eltern der Datenerhebung bei ihren Schülern zustimmen, jedoch selbst nicht motiviert sind, einen Fragebogen auszufüllen oder sich aufgrund mangelnder deutscher Sprachkenntnisse nicht dazu in der Lage sehen bzw. sich davor scheuen.

Das Ergebnis, dass lediglich in Bezug auf den Subtest TV, nicht jedoch in Bezug auf den Subtest BLK, ein Unterschied zwischen Schülern mit Deutsch als Muttersprache und Schülern mit einer anderen Muttersprache als Deutsch besteht, sollte aufgrund der eben genannten Problematik anhand einer zuverlässigeren Datenbasis repliziert werden. Lässt sich das Ergebnismuster erneut finden, könnte weiter untersucht wer-

den, worauf der Unterschied im Textverständnis zurückzuführen ist. Beispielsweise könnte die Hypothese aufgestellt werden, dass der Unterschied durch den stark kulturell beeinflussten narrativen Text zustande kommt und der Unterschied in Bezug auf den expositorischen Text deutlich geringer ausfällt.

Des Weiteren deuten die Ergebnisse der vorliegenden Arbeit darauf hin, dass ein unterdurchschnittliches Ergebnis in beiden Subtests ein Indiz für das Vorliegen einer LRS darstellen könnte. Möglicherweise könnte anhand des Subtests BLK zudem erkannt werden, ob bei Schülern mit einer anderen Muttersprache als Deutsch Defizite im Leseverständnis lediglich auf die andere Muttersprache zurückzuführen sind (dann sollte das Ergebnis im Subtest BLK unauffällig und nur das Ergebnis im Subtest TV unterdurchschnittlich sein), oder ob eine LRS vorliegt (in diesem Fall sollten die Leistungen in beiden Subtests sehr schwach ausfallen). Diese Hypothese ist ebenfalls in weiterer Forschung anhand geeigneter Stichproben zu überprüfen.

Darüber hinaus wäre es beispielsweise interessant, längsschnittlich zu prüfen, inwieweit LESEN 6-7 und LESEN 8-9 auch prädiktiv valide sind sowie, ob LESEN 6-7 und LESEN 8-9 sensitiv für Fördereffekte sind, also ob sich z. B. empirisch gezeigte Effekte von Fördermaßnahmen mit LESEN 6-7 und LESEN 8-9 replizieren lassen. Diese Aspekte der Validität sind in der vorliegenden Arbeit aufgrund des begrenzten Projektzeitrahmens nicht geprüft worden, bieten aber hinsichtlich zukünftiger Einsatzmöglichkeiten der vorliegenden Tests interessante Fragestellungen.

Kapitel 21

Fazit

Wenngleich, wie soeben dargestellt, einige Fragen offen blieben und weiterer Forschung bedürfen, und obwohl einige Kritikpunkte am methodischen Vorgehen bei der Testkonstruktion und der empirischen Erprobung von LESEN 6-7 und LESEN 8-9 nicht von der Hand zu weisen sind, sind die Ergebnisse der vorliegenden Arbeit insgesamt sehr zufriedenstellend.

Bei der Konstruktion von LESEN 6-7 und LESEN 8-9 wurde zum einen der aktuelle Forschungsstand zum Leseverständnis berücksichtigt und zum anderen wurden beide zur Verfügung stehenden Testtheorien (KTT und IRT) herangezogen. Die empirische Erprobung zeigte, dass die Tests nahezu alle gängigen Testgütekriterien erfüllen und insbesondere als reliabel und valide bezeichnet werden können. Zudem wurden die Tests an einer großen heterogenen Stichprobe normiert. In den Normstichproben liegt jedoch zugleich eine Schwäche von LESEN 6-7 und LESEN 8-9, denn trotz aufwendiger Rekrutierung zeigte sich, dass diese nicht als repräsentativ gelten können. Auch die Ergebnisse des Andersen-Tests auf Personenhomogenität sind kritisch zu sehen.

Dennoch werden mit LESEN 6-7 und LESEN 8-9 zwei Tests zur Verfügung gestellt, die es ermöglichen, einen differenzierten Eindruck vom Leseverständnis einzelner Schüler sowie ganzer Schulklassen zu bekommen. Sie können Hinweise auf das Vorliegen gravierender Leseprobleme geben und zeigen Ansatzpunkte für Fördermaßnahmen auf. Zudem sind die Tests praktisch erprobt und ökonomisch anwendbar. Die Möglichkeit eines Einsatzes als Gruppentest, macht sie nicht nur für eine Anwendung durch Lehrkräfte im Schulalltag interessant, sondern auch für Forschungszwecke. Darüber hinaus sind die Tests aber auch für die Individualdiagnostik geeignet und können daher in schulpsychologischen Beratungsstellen und Erziehungsberatungsstellen (z. B. bei einer LRS-Diagnostik) eingesetzt werden. Da LESEN 8-9 auch bei Schulabgängern der Hauptschule anwendbar ist, ergibt sich zudem die Möglichkeit, diesen Test für die Berufseignungsdiagnostik einzusetzen. Weiter ist es denkbar, LESEN 6-7 und LESEN 8-9 in Alphabetisierungskursen für funktionelle Analphabeten oder im Rahmen der Eingliederung von Menschen mit Migrationshintergrund zu verwenden.

LESEN 6-7 und LESEN 8-9 schließen somit eine bis dato bestehende Lücke in der Leseverständnisdiagnostik für die mittleren Klassenstufen und sind zurzeit für die Gruppendiagnostik in den Zielgruppen ab der siebten Klassenstufe die einzigen zur Verfügung stehenden Tests, die das Leseverständnis umfassend abbilden und im gesamten Leistungsspektrum differenzieren.

Literaturverzeichnis

Aaron, P. G., Joshi, M. & Williams, K. A. (1999). Not all reading disabilities are alike. *Journal of Learning Disabilities*, *32*, 120–137.

Adams, B. C., Bell, L. C. & Perfetti, C. A. (1995). A trading relationship between reading skill and domain knowledge in children's text comprehension. *Discourse Processes*, *20*, 307–323.

Adams, M. & Guillemin, V. (1996). *Measure theory and probability*. Boston: Birkhäuser.

Adams, M. J. (1990). *Beginning to read*. Cambridge, MA: MIT Press.

Adam-Schwebe, S., Souvignier, E., Gold, A., Hasselhorn, M., Schneider, W. & Marx, H. (2009). Der Frankfurter Leseverständnistest 5-6. In W. Schneider & W. Lenhard (Hrsg.), *Diagnostik und Förderung des Leseverständnisses* (S. 113–130). Göttingen: Hogrefe.

Adlof, S., Catts, H. & Little, T. (2006). Should the simple view of reading include a fluency component? *Reading and Writing*, *19*, 933–958.

Alba, R. & Nee, V. (1997). Rethinking assimilation theory for a new era of immigration. *International Migration Review*, *31*, 826–874.

Alexander, P. A. & Judy, J. E. (1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research*, *58*, 375–404.

Alt, C. & Bien, W. (1994). Gewichtung, ein sinnvolles Verfahren in den Sozialwissenschaften? Fragen, Probleme und Schlußfolgerungen. In S. Gabler, J. Hoffmeyer-Zlotnik & D. Krebs (Hrsg.), *Gewichtung in der Umfragepraxis* (S. 124–140). Opladen: Westdeutscher Verlag.

Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (4. Aufl.). Heidelberg: Springer.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* (5. Aufl.). Washington, DC: American Psychiatric Publishing.

Andersen, E. (1972). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123-140.

Anderson, T. H. & Armbruster, B. B. (1984). Studying. In P. D. Pearson (Hrsg.), *Handbook of Reading Research* (S. 657–679). New York: Longmann.

Andrich, D. & Godfrey, J. R. (1979). Hierarchies in the skills of Davis' Reading Comprehension Test, Form D: An empirical investigation using a latent trait model. *Reading Research Quarterly*, *14*, 182–200.

Anthoni, H., Sucheston, L., Lewis, B. A., Tapia-Páez, I., Fan, X., Zucchelli, M. et al. (2012). The aromatase gene CYP19A1: Several genetic and functional lines of evidence supporting a role in reading, speech and language. *Behavior Genetics*, *42*, 509–527.

Antoniou, F. & Souvignier, E. (2007). Strategy instruction in reading comprehension: An intervention study for students with learning disabilities. *Learning Disabilities: A Contemporary Journal*, *5*, 41–57.

Artelt, C. (2006). Lernstrategien in der Schule. In H. Mandl & H. F. Friedrich (Hrsg.), *Handbuch Lernstrategien* (S. 337–351). Göttingen: Hogrefe.

Artelt, C. (2009). Diagnostische Urteile von Lehrkräften im Bereich der Lesekompetenz. In A. Bertschi-Kaufmann & C. Rosebrock (Hrsg.), *Literalität* (S. 125–136). Weinheim: Juventa.

Artelt, C., Drechsel, B., Bos, W. & Stubbe, T. C. (2008). Lesekompetenz in PISA und PIRLS/IGLU – ein Vergleich. *Zeitschrift für Erziehungswissenschaften*, *Sonderheft*, 35–52.

Artelt, C., McElvany, N., Christmann, U., Richter, T., Groeben, N., Köster, J. et al. (2007). *Förderung von Lesekompetenz – Expertise* (BMBF, Hrsg.). Bonn: BMBF.

Artelt, C., Naumann, J. & Schneider, W. (2010). Lesemotivation und Lernstrategien. In E. Klieme et al. (Hrsg.), *PISA 2009: Bilanz nach einem Jahrzehnt* (S. 73–112). Münster: Waxmann.

Artelt, C. & Schlagmüller, M. (2004). Der Umgang mit literarischen Texten als Teilkompetenz im Lesen? Dimensionsanalysen und Ländervergleiche. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (S. 169–196). Wiesbaden: Verlag für Sozialwissenschaften.

Artelt, C., Schneider, W. & Schiefele, U. (2002). Ländervergleich zur Lesekompetenz. In J. Baumert et al. (Hrsg.), *PISA 2000 – Die Länder der Bundesrepublik Deutschland im Vergleich* (S. 55–94). Opladen: Leske + Budrich.

Artelt, C. & Stanat, P. (2010). Leistungen von Schülerinnen und Schülern im internationalen Vergleich – Die PISA Studie. In C. Spiel, B. Schober, P. Wagner & R. Reimann (Hrsg.), *Bildungspsychologie* (S. 352–356). Göttingen: Hogrefe.

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert et al. (Hrsg.), *PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 69–137). Opladen: Leske + Budrich.

Auer, M., Gruber, G., Mayringer, H. & Wimmer, H. (2005). *SLS 5-8. Salzburger Lese-Screening für die Klassenstufen 5-8*. Bern: Huber.

Bachmair, B. (2009). *Medienwissen für Pädagogen*. Wiesbaden: Verlag für Sozialwissenschaften.

Bachmann, C. (2009). *Die Flesch-Formel.* Zugriff am 12.04.2013 auf `http://www.leichtlesbar.ch/html/fleschformel.html`

Badel, I. & Valtin, R. (2005). Förderung der Lesekompetenz durch Verbesserung der Lesestrategien. In A. Sasse & R. Valtin (Hrsg.), *Lesen lehren – DGLS Beiträge* (S. 60–70). Berlin: Deutsche Gesellschaft für Lesen und Schreiben.

Baker, L. & Wigfield, A. (1999). Dimensions of children's motivation for reading and their relations to reading activities and reading achievement. *Reading Research Quarterly*, *34*, 452–477.

Barth, K. (1999). *Zur Prophylaxe von Lese- Rechtschreibstörungen: Zeitliche Verarbeitungsprozesse und ihr Zusammenhang mit phonologischer Bewußtheit und der Entwicklung von Lese- Rechtschreibkompetenz* (Dissertation, Universität Dortmund). Zugriff am 07.03.2013 auf `http://d-nb.info/960490752/34`

Bates, C. & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology*, *21*, 177–187.

Bäuerlein, K., Lenhard, W. & Schneider, W. (2012a). *LESEN 6-7. Lesetestbatterie für die Klassenstufen 6-7*. Göttingen: Hogrefe.

Bäuerlein, K., Lenhard, W. & Schneider, W. (2012b). *LESEN 8-9. Lesetestbatterie für die Klassenstufen 8-9*. Göttingen: Hogrefe.

Baumann, T., Schneider, C., Vollmar, M. & Wolters, M. (2012). *Schulen auf einen Blick* (Statistisches Bundesamt, Hrsg.). Wiesbaden: Statistisches Bundesamt. Zugriff am 07.03.2013 auf `https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Schulen/BroschuereSchulenBlick0110018129004.pdf?__blob=publicationFile`

Baumert, J. (2002). *Deutschland im internationalen Bildungsvergleich* (Bd. 57). Stuttgart: S. Hirzel.

Baumert, J. & Artelt, C. (2002). Bereichsübergreifende Perspektiven. In J. Baumert et al. (Hrsg.), *PISA 2000 – Die Länder der Bundesrepublik Deutschland im Vergleich* (S. 219–235). Opladen: Leske + Budrich.

Baumert, J. & Schümer, G. (2001). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb. In J. Baumert et al. (Hrsg.), *PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 323–407). Opladen: Leske + Budrich.

Baumert, J. & Schümer, G. (2002). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb im nationalen Vergleich. In J. Baumert et al. (Hrsg.), *PISA 2000 – Die Länder der Bundesrepublik Deutschland im Vergleich* (S. 159–202). Opladen: Leske + Budrich.

Baumert, J., Stanat, P. & Demmrich, A. (2001). PISA 2000 – Untersuchungsgegenstand, Grundlagen und Durchführung der Studie. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 15–68). Opladen: Leske + Budrich.

Baumert, J., Trautwein, U. & Artelt, C. (2003). Schulumwelten – Institutionelle Bedingungen des Lehrens und Lernens. In J. Baumert et al. (Hrsg.), *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 261–331). Opladen: Leske + Budrich.

Beck, I., McKeown, M. & Omanson, R. (1987). The effects and uses of diverse vocabulary instructional techniques. In M. G. McKeown & M. E. Curtis (Hrsg.), *The nature of vocabulary instruction* (S. 147–164). Hillsdale, N.J.: Erlbaum.

Beck, I. L., Perfetti, C. A. & McKeown, M. G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology*, *74*, 506–521.

Berger, N. (2010). *Mehr als nur ein WorT*. München: UTZ.

Böhme, K., Leucht, M., Schipolowski, S., Porsch, R., Knigge, M. & Köller, O. (2010). Anlage und Durchführung des Ländervergleichs. In O. Köller, M. Knigge & B. Tesch (Hrsg.), *Sprachliche Kompetenzen im Ländervergleich* (S. 65–85). Münster: Waxmann.

Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction*, *7*, 161–186.

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model*. Mahwah, NJ: Erlbaum.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4. überarb. Aufl.). Heidelberg: Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7., überarb. Aufl.). Berlin: Springer.

Bos, W., Lankes, E.-M., Schwippert, K., Valtin, R., Voss, A., Badel, I. & Plaßmeier, N. (2003). Lesekompetenzen deutscher Grundschülerinnen und Grundschüler am Ende der vierten Jahrgangsstufe im internationalen Vergleich. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, G. Walther & R. Valtin (Hrsg.), *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich.* Münster: Waxmann.

Brainerd, C. J. & Reyna, V. E. (1991). Fuzzy-trace theory and cognitive triage in memory development. *Developmental Psychology*, *27*, 351–369.

Brainerd, C. J. & Reyna, V. F. (1998). Fuzzy-trace theory and children's false memories. *Journal of Experimental Child Psychology*, *71*, 81–129.

Brainerd, C. J. & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science*, *11*, 164–169.

Bransford, J. D., Barclay, J. R. & Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology*, *3*, 193–209.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. akt. und erw. Aufl.). München: Pearson.

Bundesinstitut für Berufsbildung. (2010). Entwicklung und Klassifizierung der Berufe 2010 (KLdB 2010) abgeschlossen. *BWPplus. Beilage zur BWP Berufsbildung in Wissenschaft und Praxis*, *4*, 2.

Cain, K., Oakhill, J. & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, *96*, 31–42.

Carlisle, J. F. & Felbinger, L. (1991). Profiles of listening and reading comprehension. *Journal of Educational Research*, *84*, 345–354.

Carroll, J. B. (1988). The NAEP Reading Proficiency Scale is not a fiction: A reply to McLean and Goldstein. *The Phi Delta Kappan*, *69*, 761–764.

Carver, R. P. (1993). Merging the Simple View of Reading with Rauding Theory. *Journal of Literacy Research*, *25*, 439–455.

Cattinelli, I., Borghese, N. A., Gallucci, M. & Paulesu, E. (2013). Reading the reading brain: A new meta-analysis of functional imaging data on reading. *Journal of Neurolinguistics*, *26*, 214–238.

Catts, H. W. (1993). The relationship between speech-language impairments and reading disabilities. *Journal of Speech & Hearing Research*, *36*, 948–958.

Catts, H. W., Adlof, S. M. & Weismer, S. E. (2006). Language deficits in poor comprehenders: a case for the Simple View of Reading. *Journal of Speech, Language, and Hearing Research*, *49*, 278–293.

Catts, H. W. & Hogan, T. P. (2003). Language basis of reading disabilities and implications for early identification and remediation. *Reading Psychology*, *24*, 223–246.

Chi, M. T., Leeuw, N., Chiu, M.-H. & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439–477.

Chiesi, H. L., Spilich, G. J. & Voss, J. F. (1979). Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of verbal learning and verbal behavior*, *18*, 257–273.

Chiu, M. M. & McBride-Chang, C. (2006). Gender, context, and reading: A comparison of students in 43 countries. *Scientific Studies of Reading*, *10*, 331–362.

Christmann, U. (2004). Lesen. In R. Mangold, P. Vorderer & G. Bente (Hrsg.), *Lehrbuch der Medienpsychologie* (S. 419–442). Göttingen: Hogrefe.

Christmann, U. & Groeben, N. (1996). Textverstehen, Textverständlichkeit – Ein Forschungsüberblick unter Anwendungsperspektive. In H. P. Krings (Hrsg.), *Wissenschaftliche Grundlagen der technischen Kommunikation* (S. 129–189). Tübingen: Narr.

Christmann, U. & Groeben, N. (1999). Psychologie des Lesens. In B. Franzmann, K. Hasemann, D. Löffler & E. Schön (Hrsg.), *Handbuch Lesen* (S. 145–223). München: Saur.

Christmann, U. & Groeben, N. (2002). Anforderungen und Einflussfaktoren bei Sach- und Informationstexten. In N. Groeben & B. Hurrelmann (Hrsg.), *Lesekompetenz: Bedingungen, Dimensionen, Funktionen* (S. 150–173). Weinheim: Juventa.

Ciborowski, J. (1992). *Textbooks and the students who can't read them: A guide for the teaching of content*. New York: Brookline Books.

Cipielewski, J. & Stanovich, K. E. (1992). Predicting growth in reading ability from children's exposure to print. *Journal of Experimental Child Psychology*, *54*, 74–89.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum Associates.

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology*, *78*, 141–146.

Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Hrsg.), *Strategies of information processing.* London: Academic Press.

Costanzo, F., Menghini, D., Caltagirone, C., Oliveri, M. & Vicari, S. (2012). High frequency rTMS over the left parietal lobule increases non-word reading accuracy. *Neuropsychologia*, *50*, 2645–2651.

Crawford, C. B. & Koopman, P. (1979). Note: Inter-rater reliability of scree test and mean square ratio test of number of factors. *Perceptual and Motor Skills*, *49*, 223–226.

Cunningham, A. E. & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33*, 934–945.

Daneman, M. & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, *19*, 450–466.

Daneman, M. & Carpenter, P. A. (1983). Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 561–584.

Danks, J. H. & End, L. J. (1987). Processing strategies for reading and listening. In R. Horowitz & S. J. Samuels (Hrsg.), *Comprehending oral and written language* (S. 271–294). San Diego, CA: Academic Press.

de Jonge, P. & de Jong, P. F. (1996). Working memory, intelligence and reading ability in children. *Personality and Individual Differences*, *21*, 1007–1020.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods*, *2*, 292–307.

Deimel, W. (2002). Testverfahren zur Diagnostik der Lese-Rechtschreibstörung – Eine Übersicht. In G. Schulte-Körne (Hrsg.), *Legasthenie: Zum aktuellen Stand der Ursachenforschung, der diagnostischen Methoden und der Förderkonzepte* (S. 149–160). Bochum: Winkler.

de Jong, P. F. & van der Leij, A. (1999). Specific contributions of phonological abilities to early reading acquisition: Results from a dutch latent variable longitudinal study. *Journal of Educational Psychology*, *91*, 450–476.

Demaray, M. K. & Elliot, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly*, *13*, 8.

DeMars, C. (2010). *Item Response Theory. Unterstanding statistics measurement*. New York: Oxford University Press.

DESI-Konsortium. (2006). *Unterricht und Kompetenzerwerb in Deutsch und Englisch*. Frankfurt a. M.: Deutsches Institut für Internationale Pädagogische Forschung.

Diergarten, A. K. (2010). *Medien, Emotionen und Kognitionen*. Hamburg: Dr. Kovac.

Dilling, H. (Hrsg.). (2005). *Internationale Klassifikation psychischer Störungen* (5. Aufl.). Bern: Huber.

Dummer-Smoch, L. & Hackethal, R. (2011). *Kieler Leseaufbau* (8. Aufl.). Kiel: Veris.

Dutke, S. (1993). Mentale Modelle beim Erinnern sprachlich beschriebener räumlicher Anordnungen. Zur Interaktion von Gedächtnisschemata und Textrepräsentation. *Zeitschrift für experimentelle und angewandte Psychologie*, *40*, 44–71.

Dutke, S. (1994). Mentale Modelle beim Erinnern sprachlich beschriebener räumlicher Anordnungen. Zeitliche Aspekte der Modellkonstruktion und -nutzung. *Zeitschrift für experimentelle und angewandte Psychologie*, *41*, 523–548.

Dutke, S. (1996). Generic and generative knowledge: Memory schemata in the construction of mental models. In W. Battmann & S. Dutke (Hrsg.), *Processes of molar regulation of behavior* (S. 35–54). Lengerich: Pabst.

Dymock, S. (2005). Teaching expository text structure awareness. *The Reading Teacher*, *59*, 177–181.

Ehmke, T. & Jude, N. (2010). Soziale Herkunft und Kompetenzerwerb. In E. Klieme et al. (Hrsg.), *PISA 2009: Bilanz nach einem Jahrzehnt* (S. 231–254). Münster: Waxmann.

Elley, W. B. (1992). *How in the world do students read? IEA study of reading literacy* (2. Aufl.). The Hague: The International Association for the Evaluation of Educational Achievement.

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for psychologists*. New York: Psychology Press.

Ennemoser, M., Marx, P., Weber, J. & Schneider, W. (2012). Spezifische Vorläuferfertigkeiten der Lesegeschwindigkeit, des Leseverständnisses und des Rechtschreibens: Evidenz aus zwei Längsschnittstudien vom Kindergarten bis zur 4. Klasse. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *44*, 53–67.

Ennemoser, M. & Schneider, W. (2007). Relations of television viewing and reading: Findings from a 4-year longitudinal study. *Journal of Educational Psychology*, *99*, 349–368.

Esser, G. & Schmidt, M. (1993). Die langfristige Entwicklung von Kindern mit Lese-Rechtschreibschwäche. *Zeitschrift für Klinische Psychologie*, *22*, 100–116.

Esser, G., Wyschkon, A. & Schmidt, M. H. (2002). Was wird aus Achtjährigen mit einer Lese- und Rechtschreibstörung. Ergebnisse im Alter von 25 Jahren. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *31*, 235–242.

Esser, H. (1990). Familienmigration und Schulkarriere ausländischer Kinder und Jugendlicher. In H. Esser & J. Friedrichs (Hrsg.), *Generation und Identität: Theoretische und empirische Beiträge zur Migrationssoziologie* (S. 127–146). Opladen: Westdeutscher Verlag.

Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Feinberg, A. B. & Shapiro, E. S. (2003). Accuracy of teacher judgments in predicting oral reading fluency. *School Psychology Quarterly*, *18*, 52–65.

Fisseni, H.-J. (2004). *Lehrbuch der psychologischen Diagnostik* (3. überarb. Aufl.). Göttingen: Hogrefe.

Fleisher, L. S., Jenkins, J. R. & Pany, D. (2013). Effects on poor readers' comprehension of training in rapid decoding. *Reading Research Quarterly*, *15*, 30–48.

Fletcher, C. R. (1994). Levels of representation in memory for discourse. In M. A. Gernsbacher (Hrsg.), *Handbook of Psycholinguistics* (S. 589–607). San Diego, CA: Academic Press.

Florit, E. & Cain, K. (2011). The Simple View of Reading: Is it valid for different types of alphabetic orthographies? *Educational Psychology Review*, *23*, 553–576.

Friedman, N. P. & Miyake, A. (2000). Differential roles for visuospatial and verbal working memory in situation model construction. *Journal of Experimental Psychology*, *129*, 61–83.

Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. Patterson, J. Marshall & M. Coltheart (Hrsg.), *Neuropsychological and cognitive studies of phonological reading* (S. 301–330). London: Erlbaum.

Gaile, D. (2009). *Praxisbericht: „Lesen macht schlau" – Neue Lesepraxis für weiterführende Schulen*. Nürnberg: Bundesamt für Migration und Flüchtlinge.

Gaile, D. & Schoenbach, R. (2006). *Lesen macht schlau*. Berlin: Cornelsen.

Galaburda, A. M., Sherman, G. F., Rosen, G. D., Aboitiz, F. & Geschwind, N. (1985). Developmental dyslexia: Four consecutive patients with cortical anomalies. *Annals of Neurology*, *18*, 222–33.

Ganea, P. A., Shutts, K., Spelke, E. S. & DeLoache, J. S. (2007). Thinking of things unseen: Infants' use of language to update mental representations. *Psychological Science*, *18*, 734–739.

Gasteiger Klicpera, B. & Klicpera, C. (2004). Lese-Rechtschreib-Schwäche. In G. W. Lauth, M. Grünke & J. C. Brunstein (Hrsg.), *Interventionen bei Lernstörungen. Förderung, Training und Therapie in der Praxis* (S. 46–54). Göttingen: Hogrefe.

Gathercole, S. E. & Baddeley, A. D. (1993). *Working memory and language.* Lawrence Erlbaum Associates, Inc.

Gathercole, S. E., Willis, C. & Baddeley, A. D. (1991). Differentiating phonological memory and awareness of rhyme: Reading and vocabulary development in children. *British Journal of Psychology*, *82*, 387–406.

Georgiou, G. K., Das, J. P. & Hayward, D. (2009). Revisiting the „Simple View of Reading" in a group of children with poor reading comprehension. *Journal of Learning Disabilities*, *42*, 76–84.

Gleitman, L. R. & Gleitman, H. (1992). A picture is worth a thousand words, but that's the problem: The role of syntax in vocabulary acquisition. *Current Directions in Psychological Science*, *1*, 31–35.

Gold, A. (2007). Wichtige Lesestrategien. Beispiele für erfolgreiches Lernen in den Klassen 5 und 6. *Schulmagazin 5 bis 10*, *9*, 9–12.

Gold, A. (2009). Leseflüssigkeit. In A. Bertschi-Kaufmann & C. Rosebrock (Hrsg.), *Literalität* (S. 151–164). Weinheim: Juventa.

Gold, A., Mokhlesgerami, J., Rühl, K., Schreblowski, S. & Souvignier, E. (2004). *Wir werden Textdetektive – Lehrermanual & Arbeitsheft*. Göttingen: Vandenhoeck und Ruprecht.

Goldhammer, F. & Hartig, J. (2008). Interpretation von Testresultaten und Testeichung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 165–192). Heidelberg: Springer.

Goldman, S. R. & Rakestraw, J. A. (2000). Structural aspects of constructing meaning from text. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson & R. Barr (Hrsg.), *Handbook of Reading Research. Vol. III* (S. 311–335). Hillsdale, N.J.: Erlbaum.

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, N.J.: Erlbaum.

Gough, P. B., Hoover, W. A. & Peterson, C. L. (1996). Some observations on a simple view of reading. In C. Cornoldi & J. Oakhill (Hrsg.), *Reading comprehension difficulties: Processes and intervention* (S. 1–13). Mahwah NJ: Erlbaum.

Gough, P. B. & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*, 6–10.

Graesser, A. C., Louwerse, M. M., McNamara, D. S., Olney, A., Cai, Z. & Mitchell, H. H. (2007). Inference generation and cohesion in the construction of situation models: Some connections with computational linguistics. In F. Schmalhofer & C. A. Perfetti (Hrsg.), *Higher level language processes in the brain: Inference and comprehension processes* (S. 289–310). Mahwah NJ: Lawrence Earlbaum Associates.

Graesser, A. C., Millis, K. K. & Zwaan, R. A. (1997). Discoursce Comprehension. *Annual Review of Psychology*, *48*, 163–189.

Graesser, A. C., Singer, M. & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371–395.

Grigorenko, E. L. (2004). Genetic bases of developmental dyslexia: A capsule review of heritability estimates. *Enfance*, *56*, 273–288.

Grigorenko, E. L. (2011). At the junction of genomic and social sciences: An example of reading ability and disability. *Psykhe*, *20*, 3–14.

Grimm, H. (1995). Gestörter Sprachlernprozess: Ursachen und schulische Folge. In W. Niemeyer (Hrsg.), *Kommunikation und Lese-Rechtschreibschwäche* (S. 53–70). Bochum: Winkler.

Groeben, N. (1978). *Die Verständlichkeit von Unterrichtstexten*. Münster: Aschendorff.

Groeben, N. (1982). *Leserpsychologie: Textverständnis – Textverständlichkeit*. Münster: Aschendorff.

Groeben, N. (2004). Funktionen des Lesens – Normen der Gesellschaft. In N. Groeben & B. Hurrelmann (Hrsg.), *Lesesozialisation in der Mediengesellschaft* (S. 11–35). Weinheim: Juventa.

Groeben, N. & Christmann, U. (1989). Textoptimierung unter Verständlichkeitsperspektive. In G. Antos & H.-P. Krings (Hrsg.), *Textproduktion* (S. 165–196). Tübingen: Niemeyer.

Gräsel, C., Göbel, K. & Stark, R. (2007). Entwicklung von Lesekompetenz in der Sekundarstufe. In O. Böhm-Kasper, C. Schuchart & U. Schulzeck (Hrsg.), *Kontexte von Bildung. Erweiterte Perspektiven in der Bildungsforschung.* Münster: Waxmann.

Günther, K. B. (1986). Ein Stufenmodell der Entwicklung kindlicher Lese- und Schreibstrategien. In H. Brügelmann (Hrsg.), *ABC und Schriftsprache: Rätsel für Kinder, Lehrer und Forscher* (S. 32–54). Konstanz: Faude.

Guthrie, J. T., Wigfield, A., Metsala, J. L. & Cox, K. E. (1999). Motivational and cognitive predictors of text comprehension and reading amount. *Scientific Studies of Reading*, *3*, 231–256.

Hacker, W. & Osterland, D. (1995). Mentale Koordinationskapazität. Einfluß von Text- und Arbeitsgedächtnismerkmalen auf das Verstehen von Instruktionstexten. *Zeitschrift für Experimentelle Psychologie*, *42*, 646–671.

Halpern, D. F. & LaMay, M. L. (2000). The smarter sex: A critical review of sex differences in intelligence. *Educational Psychology Review*, *12*, 229–246.

Halpern, D. F. & Tan, U. (2001). Stereotypes and steroids: Using a psychobiosocial model to understand cognitive sex differences. *Brain & Cognition*, *45*, 392–414.

Harlaar, N., Kovas, Y., Dale, P. S., Petrill, S. a. & Plomin, R. (2012). Mathematics is differentially related to reading comprehension and word decoding: Evidence from a genetically sensitive design. *Journal of Educational Psychology*, *104*, 622–635.

Hasselhorn, M. (1983). Gezielte Förderung der Lernkompetenz am Beispiel der Textverarbeitung. *Unterrichtswissenschaft*, *11*, 370–382.

Hasselhorn, M. (2010). Metakognition. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (2. Aufl., S. 541–547). Weinheim: Beltz.

Hasselhorn, M. & Schuchardt, K. (2006). Lernstörungen. Eine kritische Skizze zur Epidemiologie. *Kindheit und Entwicklung*, *15*, 208–215.

Helmke, A. (2004). *Unterrichtsqualität erfassen, bewerten, verbessern*. Seelze: Kallmeyer.

Hoge, R. D. & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, *59*, 297–313.

Hoover, W. a. & Gough, P. B. (1990). The Simple View of Reading. *Reading and Writing*, *2*, 127–160.

Hopkins, K. D., George, C. A. & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement*, *22*, 177–182.

Hulslander, J., Olson, R. K., Willcutt, E. G. & Wadsworth, S. J. (2010). Longitudinal stability of reading-related skills and their prediction of reading development. *Scientific Studies of Reading*, *14*, 111 - 136.

Hunt, E., Lunneborg, C. & Lewis, J. (1975). What does it mean to be high verbal? *Cognitive Psychology*, *7*, 194–227.

Hurrelmann, B. (2002). Sozialhistorische Rahmenbedingungen von Lesekompetenz sowie soziale und personale Einflussfaktoren. In N. Groeben & B. Hurrelmann (Hrsg.), *Lesekompetenz* (S. 123–149). Weinheim: Juventa.

Hurrelmann, B. (2004). Sozialisation der Lesekompetenz. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (S. 37–60). Wiesbaden: Verlag für Sozialwissenschaften.

Hurrelmann, B. (2006). Ein erweitertes Konzept von Lesekompetenz und Konsequenzen für die Leseförderung. In G. Auernheimer (Hrsg.), *Schieflagen im Bildungssystem* (2. Aufl., S. 161–176). Wiesbaden: Verlag für Sozialwissenschaften.

Irwin, J. W. (1991). *Teaching reading comprehension processes* (2. Aufl.). New York: Pearson Education.

Ito, K. (1969). On the effect of heteroscedasticity and non-normality upon some multivariate test procedures. In P. R. Krishnaiah (Hrsg.), *Multivariate analysis* (S. 87–120). New York: Academic Press.

Ito, K. & Schul, W. J. (1964). On the robustness of the t2/0 test in multivariate analysis of variance when variance-covariance matrices are not equal. *Biometrika*, *51*, 71–82.

Jackson, M. D. & McClelland, J. L. (1979). Processing determinants of reading speed. *Journal of Experimental Psychology: General*, *108*, 151–181.

Jansen, M. G. H. (1997a). The Rasch Model for Speed Tests and Some Extensions With Applications to Incomplete Designs. *Journal of Educational and Behavioral Statistics*, *22*, 125–140.

Jansen, M. G. H. (1997b). Rasch's model for reading speed with manifest explanatory variables. *Psychometrika*, *62*, 393–409.

Jansen, M. G. H. (2003). Estimating the parameters of a structural model for the latent traits in Rasch's model for speed tfests. *Applied Psychological Measurement*, *27*, 138–151.

Jansen, M. G. H. (2007). Testing for local dependence in Rasch's multiplicative gamma model for speed tests. *Journal of Educational and Behavioral Statistics*, *32*, 24–38.

Jansen, M. G. H. & Glas, C. A. W. (2005). Checking the assumptions of Rasch's model for speed tests. *Psychometrika*, *70*, 671–684.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Johnston, T. C. & Kirby, J. R. (2006). The contribution of naming speed to the Simple View of Reading. *Reading and Writing*, *19*, 339–361.

Joshi, R. M. & Aaron, P. (2000). The Component Model of Reading: Simple View of Reading made a little more complex. *Reading Psychology*, *21*, 85–97.

Jude, N. & Klieme, E. (2010). Das Programme for International Student Assessment (PISA). In E. Klieme et al. (Hrsg.), *PISA 2009 – Bilanz nach einem Jahrzehnt* (S. 11–21). Münster: Waxmann.

Kameenui, E. J., Carnine, D. W. & Freschi, R. (1982). Effects of text construction and instructionalprocedures for teaching word meanings on comprehension and recall. *Reading Research Quarterly*, *17*, 367–388.

Kao, G. & Tienda, M. (1995). Optimism and achievement: The educational performance of immigrant youth. *Social Science Quarterly*, *76*, 1–19.

Kardash, C. M. & Almund, J. T. (1991). Self-reported learning strategies and learning from expository text. *Contemporary Educational Psychology*, *16*, 117–138.

Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen. *Zeitschrift für Pädagogische Psychologie*, *23*, 197–209.

Karing, C., Matthäi, J. & Artelt, C. (2011). Genauigkeit von Lehrerurteilen über die Lesekompetenz ihrer Schülerinnen und Schüler in der Sekundarstufe I – Eine Frage der Spezifität? *Zeitschrift für Pädagogische Psychologie*, *25*, 159–172.

Kershaw, S. & Schatschneider, C. (2012). A latent variable approach to the Simple View of Reading. *Reading and Writing*, *25*, 433–464.

Kiewra, K. A. (1989). A review of note-taking: The encoding-storage paradigm and beyond. *Educational Psychology Review*, *1*, 147–172.

King, J. & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*, 580–602.

Kintsch, E. & Kintsch, W. (1997). Learning from text. In F. E. Weinert & E. de Corte (Hrsg.), *International Encyclopedia of Developmental and Instructional Psychology.* Amsterdam: Elsevier Science.

Kintsch, W. (1994). Kognitionspsychologische Modelle des Textverstehens: Literarische Texte. In K. Reusser (Hrsg.), *Verstehen. Psychologischer Prozess und didaktische Aufgabe* (2. Aufl., S. 39–53). Bern: Huber.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY: Cambridge University Press.

Kirsch, I., de Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J. & Monseur, C. (2002). *Lesen kann die Welt verändern. Leistung und Engagement im Ländervergleich. Ergebnisse von PISA 2000*. Paris: OECD.

Kirsch, I. S., Jungeblut, A. & Mosenthal, P. B. (1998). The measurement of adult literacy. In T. S. Murray, I. S. Kirsch & L. Jenkins (Hrsg.), *Adult literacy in OECD countries: Technical report on the first International Adult Literacy Survey.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kletzien, S. B. (1991). Strategy use by good comprehenders reading expository text of differing levels. *Reading Research Quarterly*, *26*, 67–86.

Klicpera, C. & Gasteiger Klicpera, B. (1993). *Lesen und Schreiben – Entwicklung und Schwierigkeiten*. Bern: Huber.

Klicpera, C. & Gasteiger Klicpera, B. (2001). Macht Intelligenz einen Unterschied? Rechtschreiben und phonologische Fertigkeiten bei diskrepanten und nichtdiskrepanten Lese/Rechtschreibschwierigkeiten. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie*, *29*, 37–49.

Klicpera, C. & Gasteiger-Klicpera, B. (1995). *Psychologie der Lese- und Schreibschwierigkeiten. Entwicklung, Ursachen, Förderung.* Weinheim: Beltz.

Klicpera, C., Humer, R., Lugmayr, A. & Gasteiger Klicpera, B. (1993). Vorhersage von Lese- und Rechtschreibschwierigkeiten zu Beginn der 1. Klasse: Frühzeitige Differenzierung unterschiedlicher Verlaufsformen. *Frühförderung interdisziplinär*, 176–185.

Klicpera, C., Schabmann, A. & Gasteiger Klicpera, B. (1993). Lesen- und Schreibenlernen während der Pflichtschulzeit: Eine Längsschnittuntersuchung über die Häufigkeit und Stabilitat von Lese- und Rechtschreibschwierigkeiten in einem Wiener Schulbezirk. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie*, *21*, 214–225.

Klieme, E. & Stanat, P. (2002). Zur Aussagekraft internationaler Schulleistungsvergleiche: Befunde und Erklärungsansätze am Beispiel von PISA: Schulleistungen im internationalen Vergleich. *Bildung und Erziehung*, *55*, 25–44.

Kline, P. (2002). *An easy guide to factor analysis*. London: Routledge.

Knighton, T. & Bussière, P. (2006). *Educational outcomes at age 19 associated with reading ability at age 15*. Ottawa: Statistics Canada.

Köller, O., Knigge, M. & Tesch, B. (Hrsg.). (2010). *Sprachliche Kompetenzen im Ländervergleich*. Münster: Waxmann.

Körkel, J. & Hasselhorn, M. (1987). Textlernen als Problemlösen. Differentielle Aspekte und Förderperspektiven im Schulalter. In H. Neber (Hrsg.), *Angewandte Problemlösepsychologie* (S. 193–214). Münster: Aschendorff.

Köster, J. (2006). Inferenzbildung – Was Vorwissen für die Lesekompetenz bedeutet. In G. Gaiser & S. Münchenbach (Hrsg.), *Leselust dank Lesekompetenz. Leseerziehung als fächerübergreifende Aufgabe* (S. 128–137). Donauwörth: Auer.

Kubinger, K., Rasch, D. & Moder, K. (2009). Zur Legende der Voraussetzungen des t-Tests für unabhängige Stichproben. *Psychologische Rundschau*, *60*, 26-27.

Kubinger, K. D. (2000). Replik auf Jürgen Rost „Was ist aus dem Rasch-Modell geworden?“: Und für die Psychologische Diagnostik hat es doch revolutionäre Bedeutung. *Psychologische Rundschau*, *51*, 33–34.

Kubinger, K. D. (2006). *Psychologische Diagnostik*. Göttingen: Hogrefe.

Kühn, P. (2007). Evaluation rezeptiver und produktiver Wortschatzkompetenzen. Plädoyer für textuelle und konstruktivistisch angelegte Wortschatzaufgaben. In H. Willenberg (Hrsg.), *Kompetenzhandbuch für den Deutschunterricht* (S. 159–167). Hohengehren: Schneider.

Landerl, K. & Willburger, E. (2009). Der Ein-Minuten-Leseflüssigkeitstest – Ein Verfahren zur Diagnose der Leistung im Wort- und Pseudowortlesen. In W. Lenhard & W. Schneider (Hrsg.), *Diagnostik und Förderung des Leseverständnisses* (S. 65–80). Göttingen: Hogrefe.

Landerl, K. & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology*, *100*, 150–161.

Lehmann, R. & Lenkeit, J. (2008). *ELEMENT – Erhebung zum Lese- und Mathematikverständnis. Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin.* Zugriff am 06.08.2013 auf `http://www.berlin.de/imperia/md/content/sen-bildung/schulqualitaet/element6_bericht_komplett.pdf?start&ts=1210843547&file=element6_bericht_komplett.pdf`

Lehmann, R. H. (1994). Lesen Mädchen wirklich besser? Ergebnisse aus der internationalen IEA-Lesestudie. In S. Richter & H. Brügelmann (Hrsg.), *Mädchen lernen ANDERS lernen Jungen. Geschlechtsspezifische Unterschiede beim Schriftspracherwerb* (S. 99–109). Bottighofen: Libelle.

Lehmann, R. H., Peek, R., Gänsfuß, R. & Husfeldt, V. (2002). *Aspekte der Lernauslage und der Lernentwicklung – Klassenstufe 9. Ergebnisse einer Längsschnittsuntersuchung in Hamburg.* Zugriff am 07.03.2013 auf `http://bildungsserver.hamburg.de/contentblob/2815692/data/pdf-schulleistungstest-lau-9.pdf`

Lenhard, A. & Lenhard, W. (2011). *Lesbarkeitsindex (LIX).* Zugriff am 04.05.2013 auf `http://www.psychometrica.de/lix.html`

Lenhard, W. (2013). *Leseverständnis und Lesekompetenz. Grundlagen – Diagnostik – Förderung*. Stuttgart: Kohlhammer.

Lenhard, W. & Artelt, C. (2009). Komponenten des Leseverständnisses. In W. Lenhard & W. Schneider (Hrsg.), *Diagnostik und Förderung des Leseverständnisses* (S. 1–17). Göttingen: Hogrefe.

Lenhard, W., Baier, H., Lenhard, A., Schneider, W. & Hoffmann, J. (2013). *context*. Göttingen: Hogrefe.

Lenhard, W. & Schneider, W. (2006). *ELFE 1-6. Ein Leseverständnistest für Erst- bis Sechstklässler*. Göttingen: Hogrefe.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: Beltz.

Limbird, C. (2007). *Phonological processing, verbal abilities, and second language literacy development among bilingual Turkish children in Germany*. Berlin: Freie Universität.

Limbird, C. & Stanat, P. (2006). Prädiktoren von Leseverständnis bei Kindern deutscher und türkischer Herkunftssprache: Ergebnisse einer Längsschnittstudie. In A. Ittel & H. Merkens (Hrsg.), *Veränderungsmessung und Längsschnittstudien in der empirischen Erziehungswissenschaft* (S. 93–123). Wiesbaden: Verlag für Sozialwissenschaften.

Linder, M. & Grissemann, H. (2000). *Zürcher Lesetest (ZLT)* (6. Aufl.). Bern: Huber.

Lingel, K., Neuenhaus, N., Artelt, C. & Schneider, W. (2010). Metakognitives Wissen in der Sekundarstufe: Konstruktion und Evaluation domänenspezifischer Messverfahren. In E. Klieme, D. Leutner & M. Kenk (Hrsg.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. Zeitschrift für Pädagogik: Beiheft, 56.* 228–238.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lorenz, C. & Artelt, C. (2009). Fachspezifität und Stabilität diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik. *Zeitschrift für Pädagogische Psychologie*, *23*, 211–222.

MacCallum, R. C., Widaman, K. F., Zhang, S. & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84–99.

Mähler, C. & Hasselhorn, M. (2000). Lern- und Gedächtnistraining bei Kindern. In K. J. Klauer (Hrsg.), *Handbuch Kognitives Training* (S. 407–429). Göttingen: Hogrefe.

Mandl, H. & Ballstaedt, S.-P. (1982). Effects of elaboration on recall of texts. In A. Flammer & W. Kintsch (Hrsg.), *Discourse processing* (S. 482–494). Amsterdam: North-Holland Publishing Company.

Marcotte, A. M. & Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology*, *47*, 315–335.

Marcus, B. & Bühner, M. (2009). *Grundlagen der Testkonstruktion.* Studienbrief, Zugriff am 21.05.2013 auf `http://beabeablog.files.wordpress.com/2010/01/3421-studienbrief.pdf`. FernUniversität.

Mardia, K. V. (1971). The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika*, *58*, 105–121.

Marx, A. E. & Stanat, P. (2011). Reading comprehension of immigrant students in Germany: research evidence on determinants and target points for intervention. *Reading and Writing*, *25*, 1929–1945.

Marx, H. & Jungmann, T. (2000). Abhängigkeit der Entwicklung des Leseverstehens von Hörverstehen und grundlegenden Lesefertigkeiten im Grundschulalter: Eine Prüfung des Simple View of Reading-Ansatzes. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *32*, 81–93.

Marx, H. & Reinhold, B. (2010). Lese-Rechtschreibschwierigkeiten. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (4. Aufl., S. 495–507). Weinheim: Beltz.

Marx, P., Weber, J.-M. & Schneider, W. (2001). Legasthenie versus allgemeine Lese-Rechtschreibschwäche. *Zeitschrift für Pädagogische Psychologie*, *15*, 85–98.

Mayringer, H. & Wimmer, H. (2003). *Salzburger Lese-Screening für die Klassenstufen 1-4*.

McElvany, N., Becker, M. & Lüdtke, O. (2009). Die Bedeutung familiärer Merkmale für Lesekompetenz, Wortschatz, Lesemotivation und Leseverhalten. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *41*, 121–131.

McKeown, M., Beck, I., Omanson, R. & Perfetti, C. (1983). The effects of long-term vocabulary instruction on reading comprehension: A replication. *Journal of Literacy Research*, *15*, 3–18.

McNamara, D. S., Kintsch, E., Songer, N. B. & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of unterstanding in learning from text. *Cognition and Instruction*, *14*, 1–43.

Metsala, J. L. & Ehri, L. C. (1998). *Word recognition in beginning literacy*. New York: Erlbaum.

Metz, U., Marx, P., Weber, J. & Schneider, W. (2003). Overachievement im Lesen und Rechtschreiben. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *35*, 127–134.

Meyer, B. J. F. & Poon, L. W. (2001). Effects of structure strategy training and signaling on recall of text. *Journal of Educational Psychology*, *93*, 141–159.

Meyer, S. (2009). *Entwicklung und Evaluation eines Trainings zur Förderung der Lesekompetenz und Lesemotivation (LekoLemo) für die Sekundarstufe I* (Dissertation). Universität Bielefeld, Bielefeld.

Moll, K. & Landerl, K. (2010). *SLRT-II. Lese- und Rechtschreibtest*. Bern: Huber.

Möller, J. & Schiefele, U. (2004). Motivationale Grundlagen der Lesekompetenz. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (S. 101–124). Wiesbaden: Verlag für Sozialwissenschaften.

Mommers, M. J. (1987). An investigation into the relation between word recognition skills, reading comprehension and spelling skills in the first two years of primary school. *Journal of Research in Reading*, *10*, 122–143.

Moosbrugger, H. (2008a). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 215–259). Berlin: Springer.

Moosbrugger, H. (2008b). Klassische Testtheorie (KTT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 99–112). Berlin: Springer.

Moosbrugger, H. & Kelava, A. (2008). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7–26). Berlin: Springer.

Morrow, L. M. (1996). *Motivating reading and writing in diverse classrooms: Social and physical contexts in a literature-based program*. Urbana, IL: National Council of Teachers of English.

Mosenthal, P. B. & Kirsch, I. S. (1991). Toward an exlanatory model of document literacy. *Discourse Processes*, *14*, 147–180.

National Institute of Child Health and Human Development. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Reports of the subgroups*. Washington, DC: Government Printing Office.

Naumann, J., Artelt, C., Schneider, W. & Stanat, P. (2010). Lesekompetenz von PISA 2000 bis PISA 2009. In E. Klieme et al. (Hrsg.), *PISA 2009 – Bilanz nach einem Jahrzehnt* (S. 23–65). Münster: Waxmann.

Nickerson, R. S. (1981). Speech, understanding, and reading: Some differences and similarities. In O. L. V. Tzeng & H. Singer (Hrsg.), *Perception of print: Reading research in Experimental Psycholgy* (S. 257–289). Hillsdale, N.J.: Erlbaum.

Nieding, G. (2006). *Wie verstehen Kinder Texte? Die Entwicklung kognitiver Repräsentationen*. Lengerich: Pabst.

Nix, D. (2011). *Förderung der Leseflüssigkeit*. Weinheim: Juventa.

Näslund, J. C. & Schneider, W. (1996). Kindergarten letter knowledge, phonological skills, and memory processes: Relative effects on early literacy. *Journal of experimental child psychology*, *62*, 30–59.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and velicer's MAP test. *Behavior research methods, instruments, & computers: A journal of the Psychonomic Society*, *32*, 396–402.

OECD & Statistics Canada. (2000). *Literacy in the information age. Final report of the International Adult Literacy Survey*. Paris: OECD und Statistics Canada.

Palladino, P., Cornoldi, C., De Beni, R. & Pazzaglia, F. (2001). Working memory and updating processes in reading comprehension. *Memory & Cognition*, *29*, 344–354.

Paris, S. G., Wasik, B. & Turner, J. C. (1991). The development of strategic readers. In R. Barr & M. L. Kamil (Hrsg.), *Handbook of Reading Research* (S. 609–639). New York: Longmann.

Penner, Z. (2003). *Forschung für die Praxis: Neue Wege der sprachlichen Förderung von Migrantenkindern*. Berg: Kon-lab.

Pennington, B. F., Gilger, J. W., Olson, R. K. & DeFries, J. C. (1992). The external validity of age- versus IQ-discrepancy definitions of reading disability: Lessons from a twin study. *Journal of Learning Disabilities*, *25*, 562–573.

Perfetti, C. A. (1989). *There are generalized abilities and one of them is reading* (L. B. Resnick, Hrsg.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Petermann, F. & Daseking, M. (2012). *Zürcher Lesetest - II (ZLT-II)*. Bern: Huber.

Philipp, M. (2011a). Entwicklungshelfer für das Lesen? Peers und ihr längsschnittlicher Beitrag für Lesemotivation und -verhalten. In A. Ittel, H. Merkens & L. Stecher (Hrsg.), *Jahrbuch Jugendforschung*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Philipp, M. (2011b). *Lesesozialisation in Kindheit und Jugend. Lesemotivation, Leseverhalten und Lesekompetenz in Familie, Schule und Peer-Beziehungen*. Stuttgart: Kohlhammer.

Philipp, M. & Schilcher, A. (2012). *Selbstreguliertes Lesen. Ein Überblick über wirksame Förderansätze*. Seelze: Kallmeyer.

Pressley, M., Wood, E. & Woloshyn, V. (1990). Elaborative interrogation and facilitation of fact learning: Why having a knowledge base is one thing and using it is quite another. In W. Schneider & F. E. Weinert (Hrsg.), *Interactions among aptitudes, strategies, and knowledge in cognitive performance* (S. 200–211). New York: Springer.

Proyer, R., Wagner-Menghin, M. & Grafinger, G. (2010). *LEVE. Leseverständnistest für Erwachsene (Computerprogramm)*. Mödling: Schuhfried.

Prüfer, P. & Rexroth, M. (2005). Kognitive Interviews. In M. u. A. Zentrum für Umfragen (Hrsg.), *ZUMA How-to-Reihe, Nr. 15*. Mannheim: Zentrum für Umfragen, Methoden und Analysen. Zugriff am 07.03.2013 auf `http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/howto/How_to15PP_MR.pdf`

Purcell-Gates, V. (1989). Written language knowledge held by low-SES, inner city children entering kindergarten. In S. McCormick & J. Zutei (Hrsg.), *Cognitive and social perspectives for literacy research and instruction. Thirty-ninth yearbook of the National Reading Conference* (S. 95–105). Chicago: National Reading Conference.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Kopenhagen: The Danish Institute for Educational Research.

Rayner, K. & Pollatsek, A. (1989). *The psychology of reading*. London: Prentice Hall.

Reimann, G. (2009). *Moderne Eignungsbeurteilung mit der DIN 33430*. Wiesbaden: Verlag für Sozialwissenschaften.

Retelsdorf, J. & Möller, J. (2011). Entiwcklung der Leseleistung in der Sekundarstufe. In E. H. Witte & J. Doll (Hrsg.), *Sozialpsychologie, Sozialisation und Schule.* Lengerich: Pabst.

Retelsdorf, J. & Möller, J. (2008). Entwicklungen von Lesekompetenz und Lesemotivation – Schereneffekte in der Sekundarstufe? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *40*, 179–188.

Reuter-Liehr, C. (1993). Behandlung der Lese-Rechtschreibschwäche nach der Grundschulzeit: Anwendung und Überprüfung eines Konzeptes. *Zeitschrift für Kinder- und Jugendpsychiatrie*, *21*, 135–147.

Reuter-Liehr, C. (2006). *Lautgetreue Lese-Rechtschreibförderung*. Bochum: Winkler.

Reyna, V. F. & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*, 1–75.

Richter, T., Christmann, U., Hurrelmann, B. & Wilkending, G. (2002). Lesekompetenz: Prozessebenen und interindividuelle Unterschiede. In N. Groeben & B. Hurrelmann (Hrsg.), *Lesekompetenz – Bedingungen, Dimensionen, Funktionen* (S. 25–58). Weinheim: Juventa.

Rizzella, M. & O'Brien, E. J. (1996). Accessing global causes during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1208–1218.

Rjosk, C., McElvany, N., Anders, Y. & Becker, M. (2011). Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung der basalen Lesefähigkeit ihrer Schülerinnen und Schüler. *Psychologie in Erziehung und Unterricht*, *58*, 92–105.

Rosebrock, C. (2004). Informelle Sozialisationsinstanz peer group. In N. Groeben & B. Hurrelmann (Hrsg.), *Lesesozialisation in der Mediengesellschaft* (S. 250–279). Weinheim: Juventa.

Rosebrock, C. & Nix, D. (2011). *Grundlagen der Lesedidaktik und der systematischen schulischen Leseförderung* (4. Aufl.). Hohengehren: Schneider.

Rosebrock, C., Nix, D., Rieckmann, C. & Gold, A. (2011). *Leseflüssigkeit fördern: Lautleseverfahren für die Primar- und Sekundarstufe*. Seelze: Kallmeyer.

Rost, D. H. (1987). Leseverständnis oder Leseverständnisse? *Zeitschrift für Pädagogische Psychologie*, *1*, 175–196.

Rost, D. H. (1989). Reading comprehension: skill or skills? *Journal of Research in Reading*, *12*, 87–113.

Rost, D. H. (1993). Assessing different components of reading comprehension: fact or fiction? *Language Testing*, *10*, 79–92.

Rost, D. H. & Buch, S. R. (2010). Leseverständnis. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (4. Aufl., S. 507–520). Weinheim: Beltz.

Rost, D. H., Czeschlik, T. & Van der Kooij, R. (1986). Differentielle Leseverständnisdiagnostik: Wunsch oder Wirklichkeit? *Diagnostica*, *32*, 248–258.

Rost, D. H. & Hartmann, A. (1992). Lesen, Hören, Verstehen. *Zeitschrift für Psychologie*, *4*, 345–361.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2., überarb. Aufl.). Bern: Huber.

Roth, F. P., Speece, D. L. & Cooper, D. H. (2002). A longitudinal analysis of the connection between oral language and early reading. *The Journal of Educational Research*, *95*, 259–272.

Rühl, K. & Souvignier, E. (2006). *Wir werden Lesedetektive – Lehrermanual & Arbeitsheft*. Göttingen: Vandenhoeck und Ruprecht.

Rumsey, J. M., Andreason, P., Zametkin, A. J., Aquino, T., King, A. C., Hamburger, S. D. et al. (1992). Failure to activate the left temporoparietal cortex in dyslexia: An oxygen 15 positron emission tomographic study. *Archives of Neurology*, *49*, 527–534.

Rumsey, J. M., Horwitz, B., Donohue, B. C., Nace, K., Maisog, J. M. & Andreason, P. (1997). Phonological and orthographic components of word recognition. A PET-rCBF study. *Brain*, *120*, 739–759.

Rupley, W. (1997). Relationship between reading comprehension and components of word recognition: Support for developmental shifts. *Journal of research and development in education*, *30*, 255–260.

Sachs, J. (1967). Recopition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics*, *2*, 437–442.

Sachs, J. (1974). Memory in reading and listening to discourse. *Memory & Cognition*, *2*, 95–100.

Samuels, S. J. (1987). Factors that influence listening and reading comprehension. In R. Horowitz & S. J. Samuels (Hrsg.), *Comprehending oral and written language* (S. 295–325). San Diego, CA: Academic Press.

Schaffner, E., Schiefele, U., Drechsel, B. & Artelt, C. (2004). Lesekompetenz. In M. Prenzel et al. (Hrsg.), *PISA 2003 – Der Bildungsstand der Jugendlichen in Deutschland. Ergebnisse des zweiten internationalen Vergleichs* (S. 93–110). Münster: Waxmann.

Schaffner, E., Schiefele, U. & Schneider, W. (2004). Ein erweitertes Verständnis der Lesekompetenz: Die Ergebnisse des nationalen Ergänzungstests. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (S. 197–242). Wiesbaden: Verlag für Sozialwissenschaften.

Schiefele, U. (1996). *Motivation und Lernen mit Texten*. Göttingen: Hogrefe.

Schiefele, U. (2004). Förderung von Interessen. In G. W. Lauth, M. Grünke & J. C. Brunstein (Hrsg.), *Intervention bei Lernstörungen* (S. 134–144). Hogrefe.

Schipolowski, S. & Böhme, K. (2010). Ländervergleich im Fach Deutsch. In O. Köller, M. Knigge & B. Tesch (Hrsg.), *Sprachliche Kompetenzen im Ländervergleich* (S. 87–97). Münster: Waxmann.

Schlagmüller, M. & Schneider, W. (2007). *WLST 7-12. Würzburger Lesestrategie-Wissenstest für die Klassen 7-12*. Göttingen: Hogrefe.

Schlagmüller, M., Visé, M. & Schneider, W. (2001). Zur Erfassung des Gedächtniswissens bei Grundschulkindern: Konstruktionsprinzipien und empirische Bewährung der Würzburger Testbatterie zum deklarativen Metagedächtnis. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *33*, 91–102.

Schneider, W. (1996). Zum Zusammenhang zwischen Metakognition und Motivation bei Lern- und Gedächtnisvorgängen. In C. Spiel, U. Kastner-Koller & P. Deimann (Hrsg.), *Motivation und Lernen aus der Perspektive lebenslanger Entwicklung* (S. 121–133). Münster: Waxmann.

Schneider, W. (2008). Entwicklung, Diagnose und Förderung der Lesekompetenz im Kindes- und Jugendalter. In C. Fischer, F. Mönks & U. Westphal (Hrsg.), *Individuelle Förderung: Begabungen entfalten – Persönlichkeiten entwickeln* (S. 131–168). Berlin: LIT-Verlag.

Schneider, W. (2009). Diagnose basaler Lesekompetenzen in der Primar- und Sekundarstufe. In W. Lenhard & W. Schneider (Hrsg.), *Diagnostik und Förderung des Leseverständnisses* (S. 45–64). Göttingen: Hogrefe.

Schneider, W., Körkel, J. & Weinert, F. E. (1989). Domain-specific knowledge and memory performance : A comparison of high- and low-aptitude children. *Journal of Educational Psychology*, *81*, 306–312.

Schneider, W. & Küspert, P. (2003). Frühe Prävention der Lese-Rechtschreibstörungen. In W. von Suchodoletz (Hrsg.), *Therapie der Lese-Rechtschreibstörung (LRS). Traditionelle und alternative Behandlungsmethoden im Überblick* (S. 108–128). Stuttgart: Kohlhammer.

Schneider, W. & Näslund, J. C. (1999). The impact of early phonological skills on reading and spelling in school: Evidence from the Munich Longitudinal Study. In F. E. Weinert & W. Schneider (Hrsg.), *Individual development from 3 to 12: Findings from the Munich Longitudinal Study* (S. 126–147). Cambridge, UK: Cambridge University Press.

Schneider, W. & Pressley, M. (1989). *Memory development between two and twenty*. Mahwah, NJ: Erlbaum.

Schneider, W., Schlagmüller, M. & Ennemoser, M. (2007). *LGVT 6-12. Lesegeschwindigkeits- und -verständnistest für die Klassen 6 bis 12*. Göttingen: Hogrefe.

Schnotz, W. (1994). *Aufbau von Wissensstrukturen. Untersuchungen zur Kohärenzbildung beim Wissenserwerb mit Texten*. Weinheim: Psychologie Verlags Union.

Schnotz, W. & Dutke, S. (2004). Kognitionspsychologische Grundlagen der Lesekompetenz: Mehrebenenverarbeitung anhand multipler Informationsquellen. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (S. 61–99). Wiesbaden: Verlag für Sozialwissenschaften.

Schoenbach, R., Greenleaf, C., Cziko, C. & Hurwitz, L. (1999). *Reading for understanding: A guide to improving reading in middle and high school classrooms*. San Francisco: Jossey-Bass Inc.

Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts*. Frankfurt a. M.: Lang.

Schrader, F.-W. (2010). Diagnostische Kompetenz von Eltern und Lehrern. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (4. Aufl., S. 102–108). Weinheim: Beltz.

Schrader, F.-W. & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern: Komponenten und Wirkungen. *Empirische Pädagogik*, *1*, 27–52.

Schreier, M. (2004). Entwicklung von Lesekompetenz – Fördernde Einflüsse des medialen Umfeldes. In N. Groeben (Hrsg.), *Lesesozialisation in der Mediengesellschaft* (S. 402–439). Weinheim: Juventa.

Schulte-Körne, G. (2004). Lese-Rechtschreib-Störung – Symptomatik, Diagnostik, Verlauf, Ursachen und Förderung. In G. Thomé (Hrsg.), *Lese-Rechtschreib-Schwierigkeiten (LRS) und Legasthenie. Eine grundlegende Einführung* (S. 64–85). Weinheim: Beltz.

Seidenberg, M. & McClelland, J. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.

Seidenberg, M. S., Waters, G. S., Barnes, M. A. & Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behavior*, *23*, 383–404.

Seigneuric, A. & Ehrlich, M.-F. (2005). Contribution of working memory capacity to children's reading comprehension: A longitudinal investigation. *Reading and Writing*, *18*, 617–656.

Seigneuric, A., Ehrlich, M.-f., Jane, V. & Yuill, N. M. (2000). Working memory resources and children's reading comprehension. *Reading and Writing*, *13*, 81–103.

Shankweiler, D. (1999). Words to meanings. *Scientific Studies of Reading*, *3*, 113–127.

Singer, M., Graesser, A. C. & Trabasso, T. (1994). Minimal or global inference during reading. *Journal of Memory and Language*, *33*, 421–441.

Singer, M., Halldorson, M., Lear, J. C. & Andrusiak, P. (1992). Validation of causal bridging inferences. *Journal of Memory and Language*, *31*, 507–524.

Snowling, M. J., Bishop, D. V. M. & Stothard, S. E. (2000). Is preschool language impairment a risk factor for dyslexia in adolescence?. *Journal of Child Psychology and Psychiatry*, *41*, 587–600.

Souvignier, E., Hasselhorn, M., Schneider, W. & Marx, H. (2009). Effektivität von Interventionen zur Verbesserung des Leseverständnisses. In W. Lenhard & W. Schneider (Hrsg.), *Diagnostik und Förderung des Leseverständnisses* (S. 185–205). Göttingen: Hogrefe.

Souvignier, E., Küppers, J. & Gold, A. (2003). Lesestrategien im Unterricht: Einführung eines Programms zur Förderung des Textverstehens in 5. Klassen. *Unterrichtswissenschaft*, *31*, 166–183.

Souvignier, E., Trenk-Hinterberger, I., Adam-Schwebe, S. & Gold, A. (2008). *FLVT 5-6. Frankfurter Leseverständnistest für 5. und 6. Klassen*. Göttingen: Hogrefe.

Spearritt, D. (1962). *Listening comprehension: A factorial analysis*. Melbourne: Green.

Spear-Swerling, L. (2006). Children's reading comprehension and oral reading fluency in easy text. *Reading and Writing*, *19*, 199–220.

Spilich, G. J., Vesonder, G. T., Chiesi, H. L. & Voss, J. F. (1979). Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior*, *18*, 275–290.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie*, *19*, 85–95.

Spinner, K. H. (2004). Lesekomptenz in der Schule. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (S. 125–138). Wiesbaden: Verlag für Sozialwissenschaften.

Spinner, K. H. (2006). Grundlagen. In K. H. Spinner (Hrsg.), *Lesekompetenz erwerben, Literatur erfahren* (S. 7–34). Berlin: Cornelsen.

Stanat, P. (2003). Schulleistungen von Jugendlichen mit Migrationshintergrund: Differenzierung deskriptiver Befunde aus PISA und PISA-E. In J. Baumert et al. (Hrsg.), *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. S 243–260). Opladen: Leske + Budrich.

Stanat, P. & Kunter, M. (2001). Geschlechterunterschiede in Basiskompetenzen. In J. Baumert et al. (Hrsg.), *PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 249–269). Opladen: Leske + Budrich.

Stanat, P., Rauch, D. & Segeritz, M. (2010). Schülerinnen und Schüler mit Migrationshintergrund. In E. Klieme et al. (Hrsg.), *PISA 2009 – Bilanz nach einem Jahrzehnt.* (S. 200–230). Waxmann.

Stanat, P. & Schneider, W. (2004). Schwache Leser unter 15-jährigen Schülerinnen und Schülern in Deutschland: Beschreibung einer Risikogruppe. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (S. 243–273). Wiesbaden: Verlag für Sozialwissenschaften.

Stanovich, K. & Siegel, L. (1994). The phenotypic performance profile of reading-disabled children: A regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology*, *86*, 24–53.

Stanovich, K. E. (1991). Discrepancy definitions of reading disability: Has intelligence led us astray? *Reading Research Quarterly*, *26*, 7–29.

Stanovich, K. E., Siegel, L. S. & Gottardo, A. (1997). Converging evidence for phonological and surface subtypes of reading disability. *Journal of Educational Psychology*, *89*, 114–127.

Stanovich, K. E. & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, *24*, 403–433.

Sternberg, R. J. & Powell, J. S. (1983). Comprehending verbal comprehension. *American Psychologist*, *38*, 878–893.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4. Aufl.). Mahwah, NJ: Lawrence Earlbaum Associates.

Stothard, S. E. & Hulme, C. (1992). Reading comprehension difficulties in children: The role of language comprehension and working memory skills. *Reading and Writing*, *4*, 245–256.

Stothard, S. E. & Hulme, C. (1995). A comparison of phonological skills in children with reading comprehension difficulties and children with decoding difficulties. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *36*, 399–408.

Streblow, L. (2004). Zur Förderung der Lesekompetenz. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (S. 275–306). Wiesbaden: Verlag für Sozialwissenschaften.

Streblow, L., Holodynski, M. & Schiefele, U. (2007). Entwicklung eines Lesekompetenz- und Lesemotivationstrainings für die siebte Klassenstufe. *Psychologie in Erziehung und Unterricht*, *54*, 287–297.

Streblow, L. & Möller, J. (2010). Lesestrategien. In T. Hascher & B. Schmitz (Hrsg.), *Pädagogische Interventionsforschung. Theoretische Grundlagen und empirisches Handlungswissen* (S. 97–110). Weinheim: Juventa.

Streblow, L., Schiefele, U. & Riedel, S. (2012). Überprüfung des revidierten Trainings zur Förderung der Lesekompetenz und der Lesemotivation (LekoLemo) für die Sekundarstufe I. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *44*, 12–26.

Strehlow, U. & Haffner, J. (2002). Definitionsmöglichkeiten und sich daraus ergebende Häufigkeit der umschriebenen Lese- bzw. Rechtschreibstörung – Theoretische Überlegungen und empirische Befunde an einer repräsentativen Stichprobe junger Erwachsener. In W. von Suchodoletz (Hrsg.), *Welche Chancen haben Kinder mit Entwicklungsstörungen?* (S. 201–218). Göttingen: Hogrefe.

Streiner, D. L. (1998). Factors affecting reliability of interpretations of scree plots. *Psychological Reports*, *83*, 687–694.

Strobl, C. (2010). *Das Rasch-Modell. Eine verständliche Einführung für Studium und Praxis.* München: Hampp.

Südkamp, A., Kaiser, J. & Möller, J. (2012). Accuracy of Teachers' Judgments of Students' Academic Achievement: A Meta-Analysis. *Journal of Educational Psychology*, *104*, 743–762.

Südkamp, A. & Möller, J. (2009). Referenzgruppeneffekte im simulierten Klassenraum. *Zeitschrift für Pädagogische Psychologie*, *23*, 161–174.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics* (5. Aufl.; S. Hartmann, Hrsg.). Boston: Pearson.

Tardieu, H., Ehrlich, M.-F. & Gyselinck, V. (1992). Levels of representation and domain-specific knowledge in comprehension of scientific texts. *Language and Cognitive Processes*, *7*, 335–351.

Tardif, T. & Craik, F. I. M. (1989). Reading a week later: Perceptual and conceptual factors. *Journal of Memory and Language*, *28*, 107–125.

Tarelli, I., Valtin, R., Bos, W., Bremerich-Vos, A. & Schwippert, K. (2012). IGLU 2011: Wichtige Ergebnisse im Überblick. In W. Bos, I. Tarelli, A. Bremerich-Vos & K. Schwippert (Hrsg.), *IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich.* Münster: Waxmann.

Tent, L. (2006). Zensuren. In D. H. Rost (Hrsg.), *Handbuch Pädagogische Psychologie* (3. Aufl., S. 873–880). Weinheim: Beltz.

Thissen, D. & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Hrsg.), *Test scoring* (S. 73–140). Mahwah, NJ: Lawrence Erlbaum.

Thorndike, R. L. (1973). *Reading comprehension education in fifteen countries*. New York: Wiley.

Tiu, R. D., Thompson, L. A. & Lewis, B. A. (2003). The role of IQ in a Component Model of Reading. *Journal of Learning Disabilities*, *36*, 424–436.

Trabasso, T. & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language*, *24*, 612–630.

Treiman, R., Kessler, B., Zevin, J. D., Bick, S. & Davis, M. (2006). Influence of consonantal context on the reading of vowels: evidence from children. *Journal of Experimental Child Psychology*, *93*, 1–24.

Tuinman, J. J. & Brady, M. E. (1974). How does vocabulary account for variance on reading comprehension tests? A preliminary instructional analysis. In *Twentythird National Reading Conference yearbook* (S. 176–184). Clemson, SC: National Reading Conference.

Turner, R. (2002). Proficiency scales construction. In R. Adams & M. Wu (Hrsg.), *PISA 2000 – Technical report* (S. 195–216). Paris: OECD.

Unsöld, I. H. (2008). *Die Bildung von Inferenzen bei der kognitiven Verarbeitung medialer Texte*. Hamburg: Dr. Kovac.

Unterberg, D. J. (2005). *Die Entwicklung von Kindern mit LRS nach Therapie durch ein sprachsystematisches Förderkonzept. Kurz- und langfristige Wirksamkeit des Förderkonzepts nach Reuter-Liehr*. Bochum: Dr. Dieter Winkler.

van den Broek, P., Fletcher, C. R. & Risden, K. (1993). Investigations of inferential processes in reading: A theoretical and methodological integration. *Discourse Processes*, *16*, 169–180.

van den Broek, P. & Gustafson, M. (1999). Comprehension and memory for texts: Three generations of reading research. In S. R. Goldman & C. Graesser, A (Hrsg.), *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso* (S. 15–34). Mahwah, NJ: Lawrence Erlbaum.

van Dijk, T. A. (1980). *Textwissenschaft*. Tübingen: Max Niemeyer.

van Dijk, T. A. & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Van Orden, G., Pennington, B. & Stone, G. (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review*, *97*, 488–522.

van Kraayenoord, C. E. & Schneider, W. (1999). Reading achievement, metacognition, reading self-concept and interest: A study of German students in grades 3 and 4. *European Journal of Psychology of Education*, *14*, 305–324.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321–327.

Velicer, W. F., Eaton, C. A. & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Hrsg.), *Problems and solutions in human assessment* (S. 41–71). Boston: Kluwer.

Voss, A. (2006). Leseverständnis – Theorie und Empirie. *Schul-Management*, *5*, 8–10.

Voss, J. F. & Silfies, L. N. (1996). Learning from history text: The interaction of knowledge and comprehension skill with text structure. *Cognition and Instruction*, *14*, 45–68.

Walter, J. (2008). Curriculumbasiertes Messen (CBM) als lernprozessbegleitende Diagnostik: Erste deutschsprachige Ergebnisse zur Validität, Reliabilität und Veränderungssensibilität eines robusten Indikators zur Lernfortschrittsmessung beim Lesen. *Heilpädagogische Forschung*, *34*, 62–79.

Walter, J. (2013). *VSL. Verlaufsdiagnostikum sinnerfassenden Lesens*. Göttingen: Hogrefe.

Weber, J., Marx, P. & Schneider, W. (2007). Die Prävention von Lese-Rechtschreibschwierigkeiten bei Kindern mit nichtdeutscher Herkunftssprache durch ein Training der phonologischen Bewusstheit. *Zeitschrift für Pädagogische Psychologie*, *21*, 65–75.

Weis, R. & Cerankosky, B. C. (2010). Effects of video-game ownership on young boys' academic and behavioral functioning: A randomized, controlled study. *Psychological Science*, *21*, 463–70.

Weitsenfelder, L. & Hofer, S. (2012). Verstehen von Texten: Leseverständnistest für Technik-Studierende. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitsenfelder (Hrsg.), *Self-assessment: Theorie und Konzepte.* München: Pabst.

Wellman, H. M. (1983). Metamemory revisited. In M. T. H. Chi (Hrsg.), *Trends in memory development research* (S. 31–51). Basel: Karger.

Wendt, H., Grölich, C., Guill, K., Scharenberg, K. & Boss, W. (2010). Die Kompetenzen der Schülerinnen und Schüler im Leseverständnis. In W. Bos & C. Grölich (Hrsg.), *KESS 8: Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 8* (S. 21–36). Münster: Waxmann.

Weng, L.-J. & Cheng, C.-P. (2012). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement*, *65*, 697–716.

Wigfield, A. & Guthrie, J. T. (1997). Motivation for reading: An overview. *Educational Psychologist*, *32*, 57–58.

Wild, E. & Möller, J. (2009). *Pädagogische Psychologie*. Heidelberg: Springer.

Wilhelm, O. & Schulze, R. (2002). The relation of speeded and unspeeded reasoning with mental speed. *Intelligence*, *30*, 537–554.

Willburger, E., Fussenegger, B., Moll, K., Wood, G. & Landerl, K. (2008). Naming speed in dyslexia and dyscalculia. *Learning and individual differences*, *18*, 224–236.

Williams, R. W. & Bowman, M. L. (2002). Current issues on neuropsychological assessment with rural populations. In F. R. Ferraro (Hrsg.), *Minority and cross cultural aspects of neuropsychological assessment.* Lisse: Swets & Zeitlinger.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Wimmer, H. (1993). Characteristics of developmental dyslexia in a regular writing system. *Applied Psycholinguistics*, *14*, 1–33.

Wimmer, H., Hartl, M. & Moser, E. (1990). Passen „englische" Modelle des Schriftspracherwerbs auf „deutsche" Kinder? Zweifel an der Bedeutsamkeit der logographischen Stufe. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *22*, 136–154.

Wimmer, H. & Mayringer, H. (2002). Dysfluent reading in the absence of spelling difficulties: A specific disability in regular orthographies. *Journal of Educational Psychology*, *94*, 272–277.

Wimmer, H., Mayringer, H. & Landerl, K. (2000). The double-deficit hypothesis and difficulties in learning to read a regular orthography. *Journal of Educational Psychology*, *92*, 668–680.

Wolf, M. & Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. *Journal of Educational Psychology*, *91*, 415–438.

Wölfel, O., Christoph, B., Kleinert, C. & Heinert, G. (2011). Gelernt ist gelernt? Grundkompetenzen von Erwachsenen. In I. für Arbeitsmarkt-und Berufsforschung (Hrsg.), *IAB-Kurzbericht. Aktuelle Analysen aus dem Institut für Arbeitsmarkt- und Berufsforschung*. Bielefeld: Bertelsmann. Zugriff am 08.03.2013 auf `http://doku.iab.de/kurzber/2011/kb0511.pdf`

Wyschkon, A. (2011). *Repräsentativität und Umfang von Normstichproben für Leistungstests: Auswirkungen auf die Diagnostik von schwachen Leistungen und Umschriebenen Entwicklungsstörungen im Grundschulalter.* Hamburg: Kovac.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.

Zwaan, R. A. (1993). *Aspects of literacy comprehension: A cognitive approach*. Philadelphia, MA: John Benjamins.

Zwaan, R. A. & Kaschak, M. P. (2009). Language in the brain, body, and world. In P. Robbins & M. Aydede (Hrsg.), *The Cambridge handbook of situated cognition.* (S. 368–381). New York, NY: Cambridge University Press.

Zwaan, R. A., Magliano, J. P. & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 386–397.

Zwaan, R. A. & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*, 162–185.

Zwaan, R. A., Radvansky, G. A., Hilliard, A. E. & Curiel, J. M. (1998). Constructing multidimensional situation models during reading. *Scientific Studies of Reading*, *2*, 199–220.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement*, *24*, 293–308.

Zwick, W. R. & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432–442.

Anhang

Anhang A: Voruntersuchungen

Tabelle 28. Fehlerhäufigkeit und deskriptive Statistiken zur Antwortzeit (in Sekunden) und zu den LPQ-Werten, jeweils pro Item gemittelt über die Gesamtstichprobe der Voruntersuchung zum Subtest BLK ($N = 41$).

		Antwortzeit				LPQ-Wert			
Item	Fehler	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max
1	2	1.98	0.59	1.14	3.94	0.52	0.18	0.00	0.88
2	3	2.59	0.67	1.25	4.26	0.37	0.14	0.00	0.66
3	0	3.20	1.04	1.71	5.95	0.34	0.10	0.17	0.58
4	0	2.45	0.91	1.37	5.70	0.45	0.13	0.18	0.73
5	1	3.65	1.83	0.17	10.34	0.43	0.87	0.00	5.79
6	1	2.82	3.56	0.08	24.35	0.46	0.18	0.00	0.78
7	0	6.20	2.50	2.45	15.90	0.19	0.07	0.06	0.41
8	3	2.82	0.69	1.59	4.25	0.36	0.14	0.00	0.63
9	0	4.24	2.03	1.96	12.55	0.28	0.10	0.08	0.51
10	1	5.03	1.99	0.30	9.51	0.29	0.50	0.00	3.39
11	4	4.93	1.73	2.21	9.12	0.20	0.10	0.00	0.35
12	1	4.75	2.00	1.98	9.57	0.24	0.11	0.00	0.51
13	1	3.83	1.93	1.66	10.36	0.31	0.13	0.00	0.60
14	0	4.12	1.68	1.89	9.67	0.28	0.10	0.10	0.53
15	0	3.49	1.03	1.73	5.80	0.31	0.10	0.17	0.58
16	2	3.32	1.07	1.89	6.60	0.31	0.12	0.00	0.53
17	1	5.17	2.63	0.54	14.16	0.25	0.27	0.00	1.86
18	0	4.32	1.72	2.01	10.99	0.26	0.08	0.09	0.50
19	0	2.87	1.25	1.41	7.51	0.40	0.15	0.13	0.71
20	0	5.19	1.51	3.21	8.91	0.21	0.06	0.11	0.31
21	1	3.54	1.21	1.82	7.84	0.30	0.10	0.00	0.55
22	4	5.33	1.66	3.22	12.14	0.19	0.08	0.00	0.31
23	4	5.27	2.47	2.00	11.90	0.21	0.12	0.00	0.50
24	6	7.66	3.31	3.52	17.90	0.13	0.08	0.00	0.28
25	0	6.02	2.63	2.99	15.44	0.19	0.06	0.07	0.33
26	4	6.44	3.38	0.99	14.64	0.17	0.09	0.00	0.34
27	0	4.06	1.27	2.26	7.68	0.27	0.08	0.13	0.44
28	2	4.39	2.14	0.08	12.39	0.25	0.11	0.00	0.48
29	0	3.13	1.00	1.56	6.15	0.35	0.11	0.16	0.64
30	0	3.61	1.39	1.77	8.46	0.31	0.11	0.12	0.57
31	1	3.62	1.33	1.94	6.48	0.31	0.11	0.00	0.52
32	1	2.58	0.82	1.42	4.65	0.41	0.14	0.00	0.70
33	1	2.36	0.69	1.30	4.28	0.45	0.13	0.00	0.77
34	1	5.27	1.42	2.52	7.99	0.20	0.07	0.00	0.40
35	2	5.29	1.61	3.08	10.01	0.20	0.07	0.00	0.33
36	3	4.74	1.43	2.91	7.79	0.21	0.08	0.00	0.34
37	1	3.51	1.40	1.62	8.61	0.32	0.13	0.00	0.62
38	1	4.39	1.40	2.20	9.10	0.24	0.08	0.00	0.45
39	1	5.43	2.11	2.79	13.18	0.20	0.07	0.00	0.36
40	2	4.91	1.77	2.48	10.94	0.21	0.08	0.00	0.40
41	1	2.55	0.64	1.39	3.77	0.41	0.13	0.00	0.72
42	2	4.65	2.39	0.56	15.36	0.27	0.26	0.00	1.78
43	0	3.66	1.68	1.83	9.91	0.32	0.11	0.10	0.55
44	1	2.84	0.82	0.82	4.89	0.36	0.10	0.00	0.60
45	0	2.24	0.68	1.28	4.83	0.48	0.13	0.21	0.78
46	1	4.06	2.27	0.88	11.89	0.30	0.17	0.00	1.14
47	2	2.34	0.83	1.12	5.98	0.44	0.17	0.00	0.90
48	3	4.04	1.52	0.13	9.21	0.25	0.11	0.00	0.51
49	0	4.94	2.32	2.14	14.33	0.24	0.09	0.07	0.47
50	6	3.98	1.52	0.63	7.45	0.27	0.25	0.00	1.58
51	2	4.00	1.68	1.93	9.25	0.28	0.12	0.00	0.52
52	4	4.63	1.63	1.62	10.84	0.22	0.11	0.00	0.62
53	3	3.80	1.60	2.09	9.16	0.28	0.12	0.00	0.48
54	2	4.09	1.40	0.14	7.51	0.25	0.09	0.00	0.44
55	3	3.87	1.93	1.81	11.74	0.28	0.12	0.00	0.55
56	7	4.45	1.53	0.91	7.85	0.19	0.11	0.00	0.37

57	5	4.32	1.37	2.28	8.52	0.23	0.12	0.00	0.44
58	3	4.91	2.12	2.36	11.66	0.22	0.10	0.00	0.42
59	2	5.06	1.50	2.76	9.47	0.20	0.08	0.00	0.36
60	3	4.23	1.61	2.06	8.40	0.25	0.12	0.00	0.49
61	13	6.31	2.76	1.83	15.42	0.14	0.12	0.00	0.45
62	1	6.01	2.78	2.66	18.60	0.18	0.07	0.00	0.30
63	2	4.86	2.20	1.59	11.98	0.22	0.09	0.00	0.39
64	2	4.23	1.30	2.22	7.14	0.25	0.10	0.00	0.45
65	0	6.14	2.55	0.08	12.03	0.49	1.99	0.08	12.89
66	3	6.51	1.83	3.30	12.2	0.16	0.07	0.00	0.30
67	3	6.12	2.04	2.60	12.17	0.17	0.08	0.00	0.38
68	1	4.59	2.08	1.43	12.51	0.25	0.10	0.00	0.50
69	0	5.61	2.32	2.85	13.26	0.20	0.06	0.08	0.35
70	1	5.99	2.82	0.15	15.09	0.34	0.99	0.00	6.48
71	4	2.76	1.26	0.86	6.04	0.38	0.20	0.00	0.78
72	1	4.47	1.51	2.23	7.70	0.25	0.09	0.00	0.45
73	0	3.49	1.27	1.99	9.34	0.31	0.08	0.11	0.50
74	1	6.19	3.65	2.94	18.34	0.20	0.08	0.00	0.34
75	1	6.03	2.82	2.85	18.16	0.19	0.07	0.00	0.35
76	2	1.92	0.56	1.07	3.48	0.55	0.21	0.00	0.93
77	2	5.12	1.98	0.07	10.04	0.20	0.08	0.00	0.35
78	0	2.03	0.79	1.17	4.98	0.55	0.16	0.20	0.86
79	2	2.12	0.64	1.34	4.89	0.48	0.15	0.00	0.75
80	1	3.57	2.22	1.64	14.82	0.33	0.13	0.00	0.61
81	0	3.89	1.42	1.55	7.01	0.29	0.11	0.14	0.65
82	1	3.93	1.39	0.26	7.93	0.36	0.58	0.00	3.91
83	3	4.26	1.75	0.10	8.68	0.24	0.12	0.00	0.53
84	3	4.73	1.28	2.57	7.49	0.21	0.09	0.00	0.39
85	1	4.04	0.93	2.54	5.71	0.25	0.07	0.00	0.39
86	2	4.54	2.60	2.12	16.96	0.25	0.11	0.00	0.47
87	2	6.33	2.17	3.09	11.59	0.17	0.07	0.00	0.32
88	5	4.73	1.97	1.66	10.51	0.23	0.14	0.00	0.60
89	0	4.60	1.83	0.10	9.36	0.48	1.58	0.11	10.37
90	1	5.00	2.06	2.62	11.52	0.22	0.08	0.00	0.38
91	2	4.51	1.42	2.34	8.11	0.23	0.09	0.00	0.43
92	4	4.29	1.40	2.53	9.67	0.23	0.10	0.00	0.40
93	8	3.77	1.55	1.32	8.64	0.26	0.19	0.00	0.76
94	0	3.16	1.57	1.49	10.60	0.37	0.13	0.09	0.67
95	1	2.69	1.00	1.33	6.71	0.41	0.14	0.00	0.75
96	1	2.65	1.76	1.14	8.77	0.47	0.19	0.00	0.88
97	0	3.45	1.15	1.74	6.33	0.32	0.10	0.16	0.58
98	0	4.04	2.19	1.83	11.52	0.31	0.13	0.09	0.55
99	4	4.58	1.73	1.98	9.74	0.23	0.12	0.00	0.51
100	1	4.75	2.24	2.14	13.31	0.24	0.09	0.00	0.47
101	1	3.34	1.43	1.74	7.18	0.34	0.14	0.00	0.57
102	0	3.62	1.66	2.08	10.69	0.31	0.10	0.09	0.48
103	0	4.36	1.59	2.35	10.23	0.25	0.07	0.10	0.43
104	0	4.55	1.33	2.27	8.83	0.24	0.07	0.11	0.44
105	0	4.05	1.15	2.17	7.23	0.27	0.07	0.14	0.46
106	6	4.29	1.48	0.19	7.48	0.22	0.12	0.00	0.56
107	1	4.65	2.01	0.19	9.06	0.24	0.10	0.00	0.46
108	1	2.11	0.65	1.23	3.97	0.50	0.17	0.00	0.81
109	0	4.32	1.91	1.87	11.14	0.27	0.10	0.09	0.54
110	0	4.84	1.46	2.82	8.75	0.22	0.06	0.11	0.36
111	1	5.04	1.78	1.93	11.17	0.22	0.09	0.00	0.52
112	3	2.50	0.96	0.08	5.13	0.39	0.17	0.00	0.67
113	0	2.20	0.72	1.28	4.76	0.50	0.14	0.21	0.78
114	1	4.47	1.89	2.03	9.94	0.25	0.10	0.00	0.49
115	1	2.55	1.15	1.23	7.30	0.44	0.16	0.00	0.82
116	1	3.48	2.01	1.48	13.11	0.34	0.14	0.00	0.68
117	2	2.83	1.21	1.20	7.04	0.40	0.18	0.00	0.83
118	2	4.77	1.80	2.18	9.80	0.23	0.10	0.00	0.46
119	0	3.54	1.38	1.93	8.09	0.32	0.09	0.12	0.52
120	1	4.97	1.89	2.52	10.10	0.22	0.08	0.00	0.40
121	0	1.83	0.59	1.13	4.34	0.59	0.15	0.23	0.88
122	4	4.13	1.78	1.76	8.83	0.26	0.15	0.00	0.57
123	0	2.99	1.04	1.70	6.19	0.37	0.11	0.16	0.59

124	1	2.59	1.07	1.32	6.51	0.44	0.16	0.00	0.76
125	0	2.41	0.67	1.26	4.17	0.45	0.14	0.24	0.79
126	2	3.74	1.51	1.87	7.93	0.30	0.13	0.00	0.54
127	3	4.32	2.47	1.43	15.14	0.25	0.13	0.00	0.58
128	0	2.90	0.81	1.57	5.04	0.37	0.90	0.20	0.64
129	0	2.69	1.01	1.68	6.52	0.41	0.12	0.15	0.59
130	0	4.51	1.47	2.08	9.00	0.25	0.08	0.11	0.48
131	2	5.50	2.71	2.52	14.07	0.20	0.08	0.00	0.40
132	1	2.78	1.52	1.18	10.21	0.42	0.17	0.00	0.85
133	0	2.41	0.62	1.40	4.30	0.44	0.11	0.23	0.72
134	4	3.76	2.25	1.41	12.67	0.32	0.17	0.00	0.71
135	2	5.02	2.09	2.74	10.68	0.22	0.09	0.00	0.37
136	1	2.98	2.05	1.20	12.77	0.42	0.18	0.00	0.84
137	0	3.50	1.99	1.57	11.33	0.35	0.14	0.09	0.64
138	1	3.55	1.48	1.91	8.07	0.32	0.12	0.00	0.52
139	0	2.67	0.86	1.43	4.97	0.42	0.14	0.20	0.70
140	1	4.09	1.54	1.66	8.61	0.28	0.12	0.00	0.60
M	1.30	4.07	1.64	1.69	9.50	0.32	0.15	0.05	0.80

Anmerkung: Eliminierte Items sind grau markiert.

Tabelle 29. LPQ, Anteil korrekter Lösungen (in %) und Bearbeitungszeit (in Sekunden) der 100 für die Endversion ausgewählten Items, gemittelt über die Gesamtstichprobe der Voruntersuchung zum Subtest BLK (N = 41).

Item	LPQ	korrekt	Zeit	Item	LPQ	korrekt	Zeit
1	0.59	100	1.83	51	0.31	95	3.32
2	0.55	100	2.03	52	0.31	100	3.62
3	0.55	95	1.92	53	0.31	100	3.49
4	0.52	95	1.98	54	0.31	100	3.61
5	0.50	98	2.11	55	0.31	100	3.49
6	0.50	100	2.20	56	0.31	98	3.83
7	0.49	100	6.14	57	0.31	98	3.62
8	0.48	100	4.60	58	0.31	100	4.04
9	0.48	100	2.24	59	0.30	98	3.54
10	0.47	98	2.65	60	0.30	98	4.06
11	0.46	98	2.82	61	0.30	95	3.74
12	0.45	100	2.45	62	0.29	100	3.98
13	0.45	98	2.36	63	0.29	98	5.03
14	0.45	100	2.41	64	0.28	93	3.80
15	0.44	100	2.41	65	0.28	98	4.09
16	0.44	95	2.34	66	0.28	93	3.87
17	0.44	98	2.59	67	0.28	100	4.24
18	0.44	98	2.55	68	0.28	100	4.12
19	0.43	98	3.65	69	0.28	95	4.00
20	0.42	98	2.78	70	0.27	95	4.65
21	0.42	98	2.98	71	0.27	100	4.32
22	0.42	100	2.67	72	0.27	100	4.06
23	0.41	100	2.69	73	0.27	85	3.98
24	0.41	98	2.58	74	0.27	100	4.05
25	0.41	98	2.55	75	0.26	90	4.13
26	0.41	98	2.69	76	0.26	100	4.32
27	0.40	100	2.87	77	0.26	80	3.77
28	0.40	95	2.83	78	0.25	93	4.32
29	0.39	93	2.50	79	0.25	98	5.17
30	0.38	90	2.76	80	0.25	100	4.36
31	0.37	93	2.95	81	0.25	98	4.04
32	0.37	100	2.90	82	0.25	93	4.23
33	0.37	100	2.99	83	0.25	95	4.23
34	0.37	100	3.16	84	0.25	93	4.04
35	0.36	98	2.84	85	0.25	95	4.54
36	0.36	98	3.93	86	0.25	98	4.47
37	0.36	93	2.82	87	0.25	98	4.47
38	0.35	100	3.13	88	0.25	95	4.09
39	0.35	100	3.50	89	0.25	95	4.39
40	0.34	100	3.20	90	0.25	100	4.51
41	0.34	98	3.48	91	0.25	98	4.59
42	0.34	98	3.34	92	0.24	93	4.26
43	0.34	98	5.99	93	0.24	98	4.75
44	0.33	98	3.57	94	0.24	98	4.75
45	0.32	98	3.51	95	0.24	98	4.39
46	0.32	100	3.45	96	0.24	100	4.94
47	0.32	98	3.55	97	0.24	98	4.65
48	0.32	90	3.76	98	0.24	100	4.55
49	0.32	100	3.66	99	0.23	90	4.58
50	0.32	100	3.54	100	0.23	95	4.51

Tabelle 30. Ergebnis der Itemanalyse der ersten Voruntersuchung zum Subtest TV für den Text *Tiefsee* ($N = 60$).

Item	r_{it}	*SD*	SK	*p*	Kohärenzebene	Repräsentationsform
1	.16	0.34	0.24	.87	Lokal	Textbasis
2	.18	0.36	0.25	.85	Lokal	Textbasis
3	**.28**	0.39	0.36	.82	Lokal	Textbasis
4	.25	0.18	0.69	.97	Lokal	Textbasis
5	**.30**	0.42	0.36	.78	Lokal	Textbasis
6	.17	0.38	0.22	.83	Lokal	Textbasis
7	.22	0.28	0.39	.92	Lokal	Textbasis
8	.24	0.47	0.26	.68	Lokal	Textbasis
9	**.40**	0.32	0.62	.88	Lokal	Textbasis
10	.11	0.25	0.21	.93	Lokal	Textbasis
11	**.32**	0.18	0.88	.97	Lokal	Textbasis
12	**.48**	0.39	0.62	.82	Lokal	Textbasis
13	.17	0.22	0.39	.95	Lokal	Textbasis
14	.22	0.34	0.33	.87	Lokal	Textbasis
15	.22	0.28	0.39	.92	Lokal	Textbasis
16	**.27**	0.50	0.27	.58	Lokal	Situationsmodell
17	.01	0.28	0.02	.92	Lokal	Situationsmodell
18	.25	0.25	0.50	.93	Lokal	Situationsmodell
19	**.37**	0.38	0.49	.83	Lokal	Situationsmodell
20	**.30**	0.45	0.34	.73	Lokal	Situationsmodell
21	.07	0.43	0.08	.77	Global	Textbasis
22	.08	0.50	0.08	.53	Global	Textbasis
23	**.40**	0.32	0.62	.88	Global	Textbasis
24	.16	0.36	0.22	.85	Global	Textbasis
25	.18	0.18	0.51	.97	Global	Textbasis
26	.13	0.13	0.49	.98	Global	Situationsmodell
27	.15	0.36	0.21	.85	Global	Situationsmodell
28	.24	0.50	0.24	.57	Global	Situationsmodell
29	.17	0.47	0.18	.68	Global	Situationsmodell
30	.19	0.36	0.27	.85	Global	Situationsmodell
31	.16	0.50	0.16	.57	kontextualisiertes Wortverständnis	
32	**.32**	0.32	0.50	.88	kontextualisiertes Wortverständnis	
33	.12	0.49	0.13	.62	kontextualisiertes Wortverständnis	
34	.11	0.46	0.12	.30	kontextualisiertes Wortverständnis	
35	.25	0.48	0.26	.35	kontextualisiertes Wortverständnis	
36	.25	0.25	0.50	.93	Metaebene	
37	.21	0.40	0.26	.80	Metaebene	
38	**.27**	0.22	0.60	.95	Metaebene	
39	**.28**	0.13	1.10	.98	Metaebene	
40	.17	0.45	0.19	.72	Metaebene	

Anmerkung: Eliminierte Items grau; signifikante Trennschärfen fett

Tabelle 31. Ergebnis der Itemanalyse der ersten Voruntersuchung zum Subtest TV für den Text *Koboldmakis* ($N = 72$).

Item	r_{it}	*SD*	SK	*p*	Kohärenzebene	Repräsentationsform
1	.16	0.26	0.32	.93	Lokal	Textbasis
2	.23	0.20	0.56	.96	Lokal	Textbasis
3	**.27**	0.12	1.14	.99	Lokal	Textbasis
4	**.27**	0.33	0.40	.88	Lokal	Textbasis
5	**.48**	0.23	1.04	.94	Lokal	Textbasis
6	.09	0.23	0.20	.94	Lokal	Textbasis
7	**.35**	0.23	0.76	.94	Lokal	Textbasis
8	**.48**	0.35	0.69	.86	Lokal	Textbasis
9	**.30**	0.39	0.38	.82	Lokal	Textbasis
10	**.29**	0.33	0.43	.88	Lokal	Textbasis
11	**.30**	0.23	0.65	.94	Lokal	Textbasis
12	**.30**	0.36	0.41	.85	Lokal	Textbasis
13	.16	0.23	0.34	.94	Lokal	Textbasis
14	.22	0.43	0.26	.76	Lokal	Textbasis
15	**.32**	0.26	0.62	.93	Lokal	Textbasis
16	.17	0.17	0.51	.97	Lokal	Situationsmodell
17	**.27**	0.45	0.30	.72	Lokal	Situationsmodell
18	.21	0.50	0.21	.44	Lokal	Situationsmodell
19	.21	0.46	0.23	.71	Lokal	Situationsmodell
20	.10	0.50	0.10	.56	Lokal	Situationsmodell
21	**.28**	0.40	0.35	.81	Global	Textbasis
22	.15	0.49	0.16	.63	Global	Textbasis
23	.09	0.50	0.09	.43	Global	Textbasis
24	.18	0.38	0.24	.83	Global	Textbasis
25	**.50**	0.49	0.51	.63	Global	Textbasis
26	**.53**	0.30	0.88	.90	Global	Situationsmodell
27	**.33**	0.46	0.36	.31	Global	Situationsmodell
28	**.29**	0.49	0.30	.61	Global	Situationsmodell
29	**.23**	0.35	0.33	.86	Global	Situationsmodell
30	**.27**	0.41	0.33	.79	Global	Situationsmodell
31	**.31**	0.23	0.68	.94	kontextualisiertes Wortverständnis	
32	**.45**	0.50	0.45	.43	kontextualisiertes Wortverständnis	
33	.13	0.45	0.14	.72	kontextualisiertes Wortverständnis	
34	**.38**	0.38	0.51	.83	kontextualisiertes Wortverständnis	
35	.04	0.39	0.05	.18	kontextualisiertes Wortverständnis	
36	**.27**	0.28	0.48	.92	Metaebene	
37	**.34**	0.48	0.35	.65	Metaebene	
38	**.27**	0.35	0.38	.86	Metaebene	
39	**.30**	0.40	0.37	.81	Metaebene	
40	.11	0.47	0.12	.68	Metaebene	

Anmerkung: Beibehaltene Items grau; signifikante Trennschärfen fett

Tabelle 32. Ergebnis der Itemanalyse der ersten Voruntersuchung zum Subtest TV für den Text *Der geheilte Patient* ($N = 63$).

Item	r_{it}	*SD*	SK	*p*	Kohärenzebene	Repräsentationsform
1	**.36**	0.32	0.57	.89	Lokal	Textbasis
2	.13	0.21	0.30	.95	Lokal	Textbasis
3	.17	0.30	0.29	.90	Lokal	Textbasis
4	**.47**	0.38	0.62	.83	Lokal	Textbasis
5	**.27**	0.25	0.55	.94	Lokal	Textbasis
6	**.60**	0.30	1.01	.90	Lokal	Textbasis
7	.21	0.46	0.23	.70	Lokal	Textbasis
8	.14	0.30	0.23	.90	Lokal	Textbasis
9	**.41**	0.25	0.84	.94	Lokal	Textbasis
10	-.03	0.21	-0.06	.95	Lokal	Textbasis
11	**.28**	0.40	0.36	.81	Lokal	Textbasis
12	.21	0.42	0.25	.78	Lokal	Textbasis
13	.17	0.30	0.29	.90	Lokal	Textbasis
14	.17	0.18	0.47	.97	Lokal	Textbasis
15					Lokal	Textbasis
16	**.46**	0.34	0.69	.87	Lokal	Situationsmodell
17	.15	0.42	0.18	.78	Lokal	Situationsmodell
18	.24	0.18	0.69	.97	Lokal	Situationsmodell
19	**.35**	0.50	0.35	.44	Lokal	Situationsmodell
20	**.37**	0.25	0.75	.94	Lokal	Situationsmodell
21	**.35**	0.50	0.35	.48	Global	Textbasis
22	**.49**	0.48	0.51	.67	Global	Textbasis
23	**.28**	0.18	0.80	.97	Global	Textbasis
24	**.38**	0.41	0.46	.79	Global	Textbasis
25	**.32**	0.42	0.38	.22	Global	Textbasis
26	**.36**	0.18	1.02	.97	Global	Situationsmodell
27	**.43**	0.35	0.61	.86	Global	Situationsmodell
28	**.32**	0.40	0.40	.81	Global	Situationsmodell
29	**.54**	0.34	0.80	.87	Global	Situationsmodell
30	**.29**	0.35	0.41	.86	Global	Situationsmodell
31	**.38**	0.48	0.40	.67	kontextualisiertes Wortverständnis	
32	.19	0.30	0.33	.90	kontextualisiertes Wortverständnis	
33	.03	0.46	0.03	.30	kontextualisiertes Wortverständnis	
34	**.26**	0.32	0.41	.89	kontextualisiertes Wortverständnis	
35	**.30**	0.30	0.50	.90	kontextualisiertes Wortverständnis	
36	**.32**	0.43	0.37	.76	Metaebene	
37	**.29**	0.38	0.38	.83	Metaebene	
38	.06	0.42	0.07	.22	Metaebene	
39	**.29**	0.43	0.34	.76	Metaebene	
40	.12	0.46	0.13	.29	Metaebene	

Anmerkung: Beibehaltene Items grau; signifikante Trennschärfen fett

Tabelle 33. Ergebnis der Itemanalyse der ersten Voruntersuchung zum Subtest TV für den Text *Der Gescheitere* ($N = 72$).

Item	r_{it}	*SD*	SK	*p*	Kohärenzebene	Repräsentationsform
1	**.43**	0.23	0.94	.94	Lokal	Textbasis
2	.06	0.35	0.09	.86	Lokal	Textbasis
3	.23	0.40	0.29	.81	Lokal	Textbasis
4	.21	0.43	0.25	.76	Lokal	Textbasis
5	**.24**	0.20	0.60	.96	Lokal	Textbasis
6	**.34**	0.41	0.42	.79	Global	Situationsmodell
7	**.32**	0.23	0.69	.94	Lokal	Textbasis
8	.21	0.28	0.38	.92	Lokal	Textbasis
9	**.46**	0.50	0.47	.43	Lokal	Textbasis
10	**.34**	0.40	0.42	.81	Lokal	Textbasis
11	**.48**	0.38	0.64	.83	Lokal	Textbasis
12	**.62**	0.36	0.86	.85	Lokal	Textbasis
13	**.39**	0.30	0.66	.90	kontextualisiertes	Wortverständnis
14	**.48**	0.28	0.86	.92	Global	Situationsmodell
15	**.46**	0.49	0.47	.63	Lokal	Situationsmodell
16	**.53**	0.44	0.60	.74	Global	Situationsmodell
17	**.50**	0.35	0.72	.86	Global	Situationsmodell
18	.23	0.50	0.23	.57	Global	Situationsmodell
19	**.50**	0.43	0.58	.76	Global	Situationsmodell
20	**.36**	0.44	0.41	.74	Global	Situationsmodell
21	.14	0.48	0.15	.64	Global	Situationsmodell
22	-.08	0.33	-0.13	.13	Global	Situationsmodell
23	.21	0.50	0.21	.44	kontextualisiertes	Wortverständnis
24	.27	0.26	0.54	.93	kontextualisiertes	Wortverständnis
25	**.44**	0.36	0.60	.85	Lokal	Situationsmodell
26	.19	0.47	0.20	.32	Global	Situationsmodell
27	**.25**	0.43	0.29	.76	Global	Situationsmodell
28	.09	0.49	0.09	.39	Global	Situationsmodell
29	.10	0.38	0.13	.83	kontextualisiertes	Wortverständnis
30	.13	0.50	0.13	.43	kontextualisiertes	Wortverständnis?
31	**.43**	0.42	0.51	.78	kontextualisiertes	Wortverständnis
32	**.37**	0.49	0.37	.39	kontextualisiertes	Wortverständnis
33	**.50**	0.33	0.74	.88	kontextualisiertes	Wortverständnis
34	**.27**	0.28	0.48	.92	Metaebene	
35	.15	0.30	0.25	.90	Metaebene	
36	.07	0.50	0.07	.56	Metaebene	
37	-.05	0.46	-0.05	.31	Metaebene	
38	**.50**	0.44	0.56	.74	Metaebene	
39	.00	0.33	0.00	.13	Metaebene	
40	**.27**	0.43	0.31	.24	Metaebene	
41	.22	0.20	0.55	.96	Lokal	Situationsmodell
42	**.47**	0.40	0.59	.81	kontextualisiertes	Wortverständnis
43	**.34**	0.50	0.34	.53	Lokal	Situationsmodell
44	**.32**	0.44	0.36	.75	kontextualisiertes	Wortverständnis
45	**.38**	0.42	0.46	.78	Global	Situationsmodell
46	**.27**	0.36	0.37	.85	Lokal	Situationsmodell
47	.19	0.30	0.32	.90	Lokal	Textbasis
48	.03	0.50	0.03	.53	Lokal	Situationsmodell

Anmerkung: Beibehaltene Items grau; signifikante Trennschärfen fett

Tabelle 34. Ergebnis der Itemanalyse der dritten Voruntersuchung zum Subtest TV für LESEN 6-7 ($N = 113$).

Item	r_{it}	*SD*	SK	KR-20$_{ohne}$	*p*	wMNSQ	*T*	Kohärenz	Repräsentation
1	.43	0.45	0.48	.85	.72	0.92	-0.70	Lokal	Situationsmodell
2	.51	0.46	0.55	.84	.69	0.89	-1.20	Global	Situationsmodell
3	.34	0.50	0.33	.85	.49	0.99	-0.10	kontextualisiertes Wortverständnis	
4	.19	0.39	0.24	.85	.82	1.07	0.50	Global	Situationsmodell
5	.22	0.50	0.22	.85	.43	1.10	1.40	Lokal	Situationsmodell
6	.26	0.46	0.29	.85	.71	1.05	0.50	Lokal	Textbasis
7	.46	0.48	0.48	.84	.64	0.92	-0.90	Lokal	Situationsmodell
8	.23	0.47	0.24	.85	.33	1.10	1.10	Metaebene	
9	.45	0.48	0.47	.84	.65	0.94	-0.70	Lokal	Textbasis
10	.50	0.45	0.55	.84	.72	0.87	-1.20	Lokal	Situationsmodell
11	.41	0.44	0.46	.85	.74	0.93	-0.60	Lokal	Textbasis
12	.37	0.48	0.38	.85	.63	0.99	-0.10	kontextualisiertes Wortverständnis	
13	.56	0.50	0.56	.84	.52	0.83	-2.40	Lokal	Situationsmodell
14	.34	0.50	0.35	.85	.42	0.99	-0.20	Global	Situationsmodell
15	.38	0.48	0.40	.85	.34	0.97	-0.30	kontextualisiertes Wortverständnis	
16	.43	0.42	0.50	.85	.77	0.92	-0.70	Global	Situationsmodell
17	.40	0.49	0.41	.85	.61	0.95	-0.60	Lokal	Textbasis
18	.23	0.37	0.31	.85	.17	1.00	0.00	Global	Situationsmodell
19	**.07**	0.46	0.08	.85	.29	1.18	1.80	Lokal	Textbasis
20	.20	0.28	0.35	.85	.09	1.01	0.10	Lokal	Situationsmodell
21	**.06**	0.50	0.06	.85	.57	1.26	3.10	Lokal	Textbasis
22	.46	0.47	0.49	.84	.68	0.92	-0.90	kontextualisiertes Wortverständnis	
23	.38	0.32	0.60	.85	.89	0.92	-0.30	Lokal	Textbasis
24	.21	0.41	0.26	.85	.79	1.08	0.60	kontextualisiertes Wortverständnis	
25	**.14**	0.50	0.14	.85	.43	1.17	**2.20**	Lokal	Situationsmodell
26	.48	0.45	0.53	.84	.73	0.89	-1.00	Global	Situationsmodell
27	.38	0.49	0.39	.85	.39	0.96	-0.50	Global	Situationsmodell
28	.30	0.38	0.39	.85	.82	1.03	0.20	Global	Situationsmodell
29	.16	0.47	0.17	.85	.67	1.12	1.30	Lokal	Textbasis
30	.37	0.42	0.44	.85	.77	0.96	-0.30	Lokal	Textbasis
31	.32	0.47	0.34	.85	.68	1.02	0.20	Global	Situationsmodell
32	**.01**	0.46	0.01	.86	.70	1.27	**2.50**	Global	Textbasis
33	.29	0.50	0.29	.85	.57	1.07	0.90	Metaebene	
34	.48	0.46	0.52	.84	.70	0.92	-0.80	Lokal	Textbasis
35	.53	0.50	0.53	.84	.48	0.85	-2.20	Global	Situationsmodell
36	.34	0.46	0.37	.85	.69	1.00	0.10	kontextualisiertes Wortverständnis	
37	.38	0.49	0.39	.85	.61	0.98	-0.20	Global	Situationsmodell
38	.20	0.50	0.20	.85	.56	1.14	1.80	Lokal	Situationsmodell
39	.38	0.47	0.41	.85	.68	0.96	-0.40	Lokal	Situationsmodell
40	.23	0.44	0.27	.85	.75	1.07	0.60	Global	Situationsmodell

Anmerkungen:
$KR-20_{ohne}$: KR-20, wenn das Item eliminiert wird
T: Teststatistik zur Signifikanzprüfung der wMNSQ
fett: nicht signifikante Trennschärfen und *T*-Werte > 2
grau: eliminierte Items

Tabelle 35. Ergebnis der Itemanalyse der dritten Voruntersuchung zum Subtest TV für LESEN 8-9 ($N = 189$).

Item	r_{it}	*SD*	SK	KR-20$_{ohne}$	*p*	wMNSQ	*T*	Kohärenz	Repräsentation
1	.32	0.35	0.46	.85	.86	1.02	0.20	Lokal	Textbasis
2	.29	0.39	0.36	.84	.81	1.11	0.90	Lokal	Textbasis
3	.43	0.38	0.57	.85	.83	0.93	-0.50	Lokal	Textbasis
4	.38	0.50	0.38	.85	.43	1.04	0.60	Lokal	Situationsmodell
5	.58	0.46	0.64	.85	.70	0.83	-2.10		Metaebene
6	.39	0.41	0.48	.85	.79	0.99	-0.10	Lokal	Situationsmodell
7	.56	0.41	0.68	.84	.78	0.84	-1.60	kontextualisiertes Wortverständnis	
8	**.16**	0.43	0.19	.85	.75	1.25	**2.50**	Lokal	Textbasis
9	.47	0.50	0.47	.84	.58	0.96	-0.50	Lokal	Textbasis
10	.35	0.45	0.38	.84	.71	1.06	0.70		Metaebene
11	.44	0.46	0.48	.85	.70	0.98	-0.20	Lokal	Situationsmodell
12	.41	0.44	0.47	.85	.74	0.98	-0.20	Global	Situationsmodell
13	.25	0.48	0.27	.84	.65	1.18	**2.30**		Metaebene
14	.43	0.50	0.43	.85	.52	1.00	0.00	Global	Situationsmodell
15	.51	0.50	0.51	.85	.55	0.92	-1.10	Global	Situationsmodell
16	.51	0.50	0.52	.85	.56	0.93	-1.10	kontextualisiertes Wortverständnis	
17	.38	0.48	0.39	.85	.37	1.01	0.20	Global	Situationsmodell
18	.39	0.49	0.40	.85	.41	1.02	0.30	Global	Situationsmodell
19	.28	0.48	0.29	.85	.37	1.12	1.70	Lokal	Situationsmodell
20	.50	0.50	0.51	.85	.42	0.89	-1.60	Lokal	Situationsmodell
21	.54	0.49	0.55	.85	.60	0.89	-1.60	Global	Situationsmodell
22	.51	0.35	0.73	.84	.86	0.81	-1.40	Lokal	Situationsmodell
23	.29	0.32	0.46	.85	.89	0.99	0.00	Global	Situationsmodell
24	.44	0.39	0.57	.85	.82	0.95	-0.40	Lokal	Textbasis
25	.40	0.44	0.46	.85	.75	0.98	-0.20	Lokal	Textbasis
26	.53	0.46	0.57	.84	.70	0.89	-1.40	kontextualisiertes Wortverständnis	
27	.42	0.50	0.42	.85	.51	1.01	0.20	Global	Situationsmodell
28	.38	0.45	0.43	.85	.72	1.06	0.70	kontextualisiertes Wortverständnis	
29	.44	0.42	0.53	.85	.78	0.92	-0.80	Global	Situationsmodell
30	.56	0.34	0.83	.85	.87	0.77	-1.70	Global	Situationsmodell
31	.41	0.49	0.42	.85	.59	1.02	0.30	kontextualisiertes Wortverständnis	
32	.26	0.50	0.26	.86	.53	1.20	**2.80**	Global	Situationsmodell
33	.41	0.49	0.42	.85	.61	1.02	0.30	Lokal	Situationsmodell
34	.38	0.49	0.39	.84	.59	1.06	0.90		Metaebene
35	.59	0.48	0.61	.84	.63	0.84	-2.30	Lokal	Situationsmodell
36	.55	0.48	0.57	.85	.64	0.88	-1.70	Lokal	Situationsmodell
37	.39	0.47	0.41	.85	.33	0.96	-0.50	Lokal	Textbasis
38	.36	0.49	0.37	.85	.39	1.07	1.00	kontextualisiertes Wortverständnis	
39	**.13**	0.50	0.13	.85	.54	1.33	**4.50**	Lokal	Situationsmodell
40	.36	0.50	0.36	.85	.52	1.10	1.40	Lokal	Textbasis

Anmerkungen:
$KR-20_{ohne}$: KR-20, wenn das Item eliminiert wird
T: Teststatistik zur Signifikanzprüfung der wMNSQ
fett: nicht signifikante Trennschärfen und *T*-Werte > 2
grau: eliminierte Items

Tabelle 36. Itemkennwerte für LESEN 6-7 ($N = 113$) nach der Itemselektion der dritten Voruntersuchung zum Subtest TV.

Item	r_{it}	*SD*	SK	$KR-20_{ohne}$	*p*	wMNSQ	*T*
1	.44	0.45	0.48	.86	.72	0.93	-0.60
2	.50	0.46	0.54	.85	.69	0.89	-1.20
3	.37	0.50	0.36	.86	.49	1.04	0.50
4	.20	0.39	0.25	.86	.82	1.09	0.60
5	.21	0.50	0.21	.86	.43	1.17	2.00
6	.26	0.46	0.28	.86	.71	1.09	0.80
7	.48	0.48	0.50	.86	.64	0.93	-0.80
8	.45	0.48	0.47	.86	.65	0.95	-0.60
9	.48	0.45	0.53	.86	.72	0.90	-1.00
10	.40	0.44	0.45	.86	.74	0.96	-0.30
11	.38	0.48	0.39	.86	.63	1.02	0.30
12	.55	0.50	0.54	.85	.52	0.86	-1.80
13	.34	0.50	0.34	.86	.42	1.03	0.30
14	.39	0.48	0.41	.86	.34	1.00	0.10
15	.44	0.42	0.52	.86	.77	0.90	-0.80
16	.38	0.49	0.38	.86	.61	1.00	0.00
17	.24	0.37	0.32	.86	.17	1.03	0.20
18	.45	0.47	0.48	.86	.68	0.92	-0.80
19	.39	0.32	0.60	.86	.89	0.91	-0.40
20	.22	0.41	0.27	.86	.79	1.11	0.80
21	.49	0.45	0.55	.86	.73	0.89	-1.10
22	.39	0.49	0.40	.86	.39	0.99	-0.10
23	.30	0.38	0.39	.86	.82	1.02	0.20
24	.16	0.47	0.17	.86	.67	1.19	2.00
25	.35	0.42	0.42	.86	.77	1.01	0.10
26	.33	0.47	0.35	.86	.68	1.04	0.50
27	.28	0.50	0.28	.86	.57	1.11	1.30
28	.46	0.46	0.50	.86	.70	0.91	-0.90
29	.52	0.50	0.52	.85	.48	0.89	-1.50
30	.32	0.46	0.35	.86	.69	1.06	0.60
31	.40	0.49	0.40	.86	.61	1.00	0.00
32	.23	0.50	0.23	.86	.56	1.15	1.80
33	.39	0.47	0.42	.86	.68	0.98	-0.10
34	.25	0.44	0.28	.86	.75	1.10	0.90

Anmerkungen:
$KR-20_{ohne}$: KR-20, wenn das Item eliminiert wird
T: Teststatistik zur Signifikanzprüfung der wMNSQ

Tabelle 37. Itemkennwerte für LESEN 8-9 ($N = 189$) nach der Itemselektion der dritten Voruntersuchung zum Subtest TV.

Item	r_{it}	*SD*	SK	$KR-20_{ohne}$	*p*	wMNSQ	*T*
1	.33	0.35	0.47	.90	.86	1.01	0.10
2	.29	0.39	0.37	.90	.81	1.12	1.00
3	.43	0.38	0.57	.90	.83	0.91	-0.70
4	.39	0.50	0.40	.90	.43	1.03	0.40
5	.58	0.46	0.64	.89	.70	0.84	-2.00
6	.40	0.41	0.49	.90	.79	0.99	0.00
7	.55	0.41	0.67	.89	.78	0.85	-1.50
8	.16	0.43	0.18	.90	.75	1.26	2.60
9	.46	0.50	0.46	.90	.58	0.99	-0.10
10	.35	0.45	0.39	.90	.71	1.08	0.90
11	.42	0.44	0.47	.90	.74	1.00	0.00
12	.25	0.48	0.26	.90	.65	1.21	2.60
13	.43	0.50	0.43	.90	.52	1.01	0.10
14	.51	0.50	0.51	.90	.55	0.92	-1.10
15	.51	0.50	0.51	.90	.56	0.93	-1.00
16	.37	0.48	0.39	.90	.37	1.03	0.40
17	.38	0.49	0.39	.90	.41	1.04	0.50
18	.28	0.48	0.29	.90	.37	1.13	1.70
19	.50	0.50	0.50	.90	.42	0.93	-1.10
20	.54	0.49	0.55	.89	.60	0.92	-1.20
21	.51	0.35	0.73	.90	.86	0.83	-1.30
22	.29	0.32	0.46	.90	.89	1.00	0.10
23	.44	0.39	0.56	.90	.82	0.96	-0.30
24	.40	0.44	0.46	.90	.75	1.00	0.00
25	.52	0.46	0.57	.90	.70	0.90	-1.30
26	.43	0.50	0.43	.90	.51	1.00	0.00
27	.38	0.45	0.43	.90	.72	1.07	0.70
28	.44	0.42	0.53	.90	.78	0.93	-0.60
29	.57	0.34	0.83	.90	.87	0.76	-1.80
30	.41	0.49	0.42	.90	.59	1.03	0.40
31	.25	0.50	0.25	.90	.53	1.22	3.00
32	.41	0.49	0.42	.90	.61	1.05	0.70
33	.38	0.49	0.39	.90	.59	1.08	1.10
34	.58	0.48	0.60	.89	.63	0.85	-2.10
35	.54	0.48	0.56	.89	.64	0.90	-1.40
36	.39	0.47	0.41	.90	.33	0.97	-0.40
37	.37	0.49	0.37	.90	.39	1.08	1.20
38	.35	0.50	0.35	.90	.52	1.12	1.80

Anmerkungen:
$KR-20_{ohne}$: KR-20, wenn das Item eliminiert wird
T: Teststatistik zur Signifikanzprüfung der wMNSQ

Tabelle 38. Ergebnisse der MAP-Tests im Rahmen der dritten Voruntersuchung zum Subtest TV für beide Tests.

	Lesen 6-7				Lesen 6-7		
Faktor	Eigenwert	Anzahl auspartialisierte Komponenten	Vierte Potenz der mittleren Partialkorrelation	Faktor	Eigenwert	Anzahl auspartialisierte Komponenten	Vierte Potenz der mittleren Partialkorrelation
1	6.46	0	0.0024	1	8.50	0	0.0036
2	2.00	1	0.0004	2	2.00	1	0.0002
3	1.76	2	0.0004	3	1.52	2	0.0002
4	1.68	3	0.0005	4	1.45	3	0.0002
5	1.61	4	0.0007	5	1.40	4	0.0003
6	1.55	5	0.0007	6	1.37	5	0.0003
7	1.40	6	0.0008	7	1.24	6	0.0003
8	1.26	7	0.0009	8	1.18	7	0.0004
9	1.17	8	0.0009	9	1.15	8	0.0005
10	1.12	9	0.0011	10	1.10	9	0.0005
11	1.06	10	0.0014	11	1.05	10	0.0006
12	1.05	11	0.0016	12	0.98	11	0.0008
13	0.98	12	0.0020	13	0.96	12	0.0009
14	0.95	13	0.0023	14	0.91	13	0.0012
15	0.87	14	0.0027	15	0.87	14	0.0015
16	0.83	15	0.0032	16	0.85	15	0.0018
17	0.81	16	0.0037	17	0.84	16	0.0020
18	0.78	17	0.0043	18	0.80	17	0.0023
19	0.67	18	0.0051	19	0.76	18	0.0028
20	0.62	19	0.0066	20	0.74	19	0.0034
21	0.60	20	0.0084	21	0.70	20	0.0045
22	0.57	21	0.0099	22	0.69	21	0.0047
23	0.54	22	0.0116	23	0.64	22	0.0056
24	0.50	23	0.0148	24	0.62	23	0.0066
25	0.47	24	0.0169	25	0.58	24	0.0078
26	0.43	25	0.0225	26	0.56	25	0.0094
27	0.41	26	0.0293	27	0.53	26	0.0117
28	0.36	27	0.0393	28	0.48	27	0.0148
29	0.34	28	0.0538	29	0.47	28	0.0178
30	0.28	29	0.0780	30	0.42	29	0.0232
31	0.26	30	0.1159	31	0.41	30	0.0308
32	0.24	31	0.1918	32	0.41	31	0.0396
33	0.18	32	0.3589	33	0.36	32	0.0533
34	0.16	33	1.0000	34	0.33	33	0.0753
–	–	–	–	35	0.31	34	0.1172
–	–	–	–	36	0.29	35	0.1916
–	–	–	–	37	0.28	36	0.3641
–	–	–	–	38	0.25	37	1.0000

Tabelle 39. Eigenwerte und Faktorladungen der Ein-Faktorenlösung für den Subtest TV im Rahmen der dritten Voruntersuchung.

LESEN 6-7					LESEN 8-9				
Faktor	Eigenwert	Item	λ	h^2	Faktor	Eigenwert	Item	λ	h^2
1	10.16	1	0.61	0.37	1	13.78	1	0.53	0.28
2	2.65	2	0.68	0.46	2	2.39	2	0.45	0.20
3	2.36	3	0.47	0.22	3	1.83	3	0.66	0.43
4	2.12	4	0.32	0.10	4	1.73	4	0.54	0.29
5	2.08	5	0.31	0.09	5	1.71	5	0.80	0.63
6	1.90	6	0.38	0.14	6	1.63	6	0.60	0.36
7	1.66	7	0.67	0.44	7	1.40	7	0.79	0.62
8	1.43	8	0.63	0.39	8	1.31	8	0.23	0.05
9	1.28	9	0.65	0.43	9	1.24	9	0.60	0.36
10	1.17	10	0.65	0.43	10	1.18	10	0.47	0.22
11	1.10	11	0.50	0.25	11	0.98	11	0.59	0.34
12	1.08	12	0.74	0.55	12	0.97	12	0.33	0.11
13	0.97	13	0.49	0.24	13	0.93	13	0.56	0.31
14	0.91	14	0.64	0.41	14	0.83	14	0.67	0.45
15	0.80	15	0.67	0.45	15	0.79	15	0.67	0.44
16	0.70	16	0.51	0.26	16	0.74	16	0.52	0.27
17	0.69	17	0.55	0.30	17	0.68	17	0.51	0.26
18	0.64	18	0.64	0.41	18	0.66	18	0.41	0.17
19	0.46	19	0.75	0.56	19	0.60	19	0.68	0.46
20	0.39	20	0.36	0.13	20	0.54	20	0.70	0.48
21	0.36	21	0.68	0.47	21	0.50	21	0.80	0.64
22	0.28	22	0.55	0.30	22	0.47	22	0.53	0.28
23	0.24	23	0.50	0.25	23	0.42	23	0.65	0.42
24	0.16	24	0.24	0.06	24	0.37	24	0.59	0.35
25	0.12	25	0.55	0.30	25	0.29	25	0.70	0.49
26	0.07	26	0.45	0.20	26	0.27	26	0.59	0.35
27	-0.02	27	0.38	0.14	27	0.24	27	0.53	0.28
28	-0.08	28	0.64	0.41	28	0.19	28	0.64	0.41
29	-0.09	29	0.70	0.49	29	0.11	29	0.89	0.79
30	-0.17	30	0.46	0.21	30	0.09	30	0.53	0.28
31	-0.31	31	0.53	0.28	31	0.03	31	0.34	0.11
32	-0.33	32	0.30	0.09	32	-0.01	32	0.56	0.32
33	-0.38	33	0.57	0.33	33	-0.06	33	0.50	0.25
34	-0.40	34	0.36	0.13	34	-0.10	34	0.77	0.59
–	–	–	–	–	35	-0.14	35	0.72	0.52
–	–	–	–	–	36	-0.18	36	0.55	0.30
–	–	–	–	–	37	-0.19	37	0.51	0.26
–	–	–	–	–	38	-0.23	38	0.51	0.26

Tabelle 40. Zuordnung der Items des Subtests TV zu den Ebenen des Leseverständnisses bzw. den Formen der Textrepräsentation für LESEN 6-7 und LESEN 8-9

LESEN 6-7

	Expositorischer Text		Narrativer Text	
Item	Kohärenz	Repräsentation	Kohärenz	Repräsentation
1	Global	Situationsmodell	Lokal	Textbasis
2	Global	Situationsmodell	Wortverständnis	
3	Lokal	Textbasis	Global	Situationsmodell
4	Lokal	Textbasis	Lokal	Textbasis
5	Lokal	Textbasis	Lokal	Situationsmodell
6	Global	Situationsmodell	Wortverständnis	
7	Lokal	Situationsmodell	Global	Situationsmodell
8	Lokal	Textbasis	Lokal	Textbasis
9	Lokal	Textbasis	Global	Situationsmodell
10	Wortverständnis		Global	Situationsmodell
11	Lokal	Textbasis	Global	Situationsmodell
12	Lokal	Situationsmodell	Lokal	Situationsmodell
13	Wortverständnis		Global	Situationsmodell
14	Global	Situationsmodell	Global	Situationsmodell
15	Global	Situationsmodell	Metaebenen	
16	Wortverständnis		Global	Situationsmodell
17	Global	Situationsmodell	Global	Situationsmodell

LESEN 8-9

	Expositorischer Text		Narrativer Text	
Item	Kohärenz	Repräsentation	Kohärenz	Repräsentation
1	Lokal	Textbasis	Lokal	Situationsmodell
2	Lokal	Textbasis	Lokal	Situationsmodell
3	Metaebenen		Lokal	Situationsmodell
4	Wortverständnis		Lokal	Situationsmodell
5	Lokal	Situationsmodell	Lokal	Textbasis
6	Metaebenen		Wortverständnis	
7	Lokal	Situationsmodell	Lokal	Situationsmodell
8	Lokal	Textbasis	Wortverständnis	
9	Lokal	Situationsmodell	Lokal	Situationsmodell
10	Metaebenen		Lokal	Situationsmodell
11	Lokal	Textbasis	Global	Situationsmodell
12	Wortverständnis		Lokal	Situationsmodell
13	Global	Situationsmodell	Wortverständnis	
14	Global	Situationsmodell	Global	Situationsmodell
15	Global	Situationsmodell	Metaebenen	
16	Global	Situationsmodell	Lokal	Situationsmodell
17	Lokal	Situationsmodell	Lokal	Situationsmodell
18	Lokal	Situationsmodell	Wortverständnis	
19	Global	Situationsmodell	Lokal	Situationsmodell

Anhang B: Normierung

Tabelle 41. Deskriptive Statistik für beide Subtests und das Gesamtergebnis der Normdaten nach Klassenstufe, Schulart und Bundesland getrennt.

		Bundes-		BLK		TV	
Klasse	Schulart	land	*N*	*M*	*SD*	*M*	*SD*
6	HS	BY	192	43.70	9.40	14.11	5.70
		BW	44	43.60	8.10	11.30	4.86
		NRW	27	39.30	6.20	11.37	5.94
	RS	BW	121	54.00	11.20	19.30	5.20
		NRW	29	44.60	12.90	16.03	5.77
	GYM	BW	140	63.00	15.50	23.00	5.25
		RLP	39	66.00	12.20	27.15	4.08
		SH	112	57.70	10.30	23.52	5.01
7	HS	BY	209	49.01	11.23	17.56	5.50
		BW	36	46.83	9.69	12.89	5.12
		NRW	37	42.03	8.58	13.51	5.40
	RS	BW	211	57.45	11.75	21.80	5.47
		NRW	41	46.61	11.05	18.17	5.17
		NS	51	54.53	16.20	18.14	5.35
	GYM	BW	57	67.68	11.41	26.70	4.08
		RLP	37	68.78	17.43	23.51	5.40
		BB	95	62.40	14.44	26.03	5.38
		SH	77	61.97	12.09	25.94	4.75
8	HS	BY	172	52.71	13.32	17.08	5.05
		NRW	42	46.60	13.02	15.81	5.64
	RS	BW	55	61.71	9.59	24.35	5.75
		B	62	61.69	17.53	18.32	6.97
		NS	43	62.26	18.95	18.79	6.13
	GYM	BW	70	71.86	9.64	28.04	4.51
		BB	86	66.00	18.70	24.94	6.24
9	HS	BY	46	58.54	12.39	19.65	5.02
		NRW	46	53.39	13.96	16.28	5.85
	RS	BW	53	66.81	13.19	25.26	5.34
		B	64	56.20	9.32	21.25	6.73
		NS	45	67.64	17.97	22.38	5.64
	GYM	BW	51	75.61	10.11	30.24	4.15
		RLP	11	73.55	11.51	28.91	4.25
		BB	86	62.92	11.99	26.52	5.78

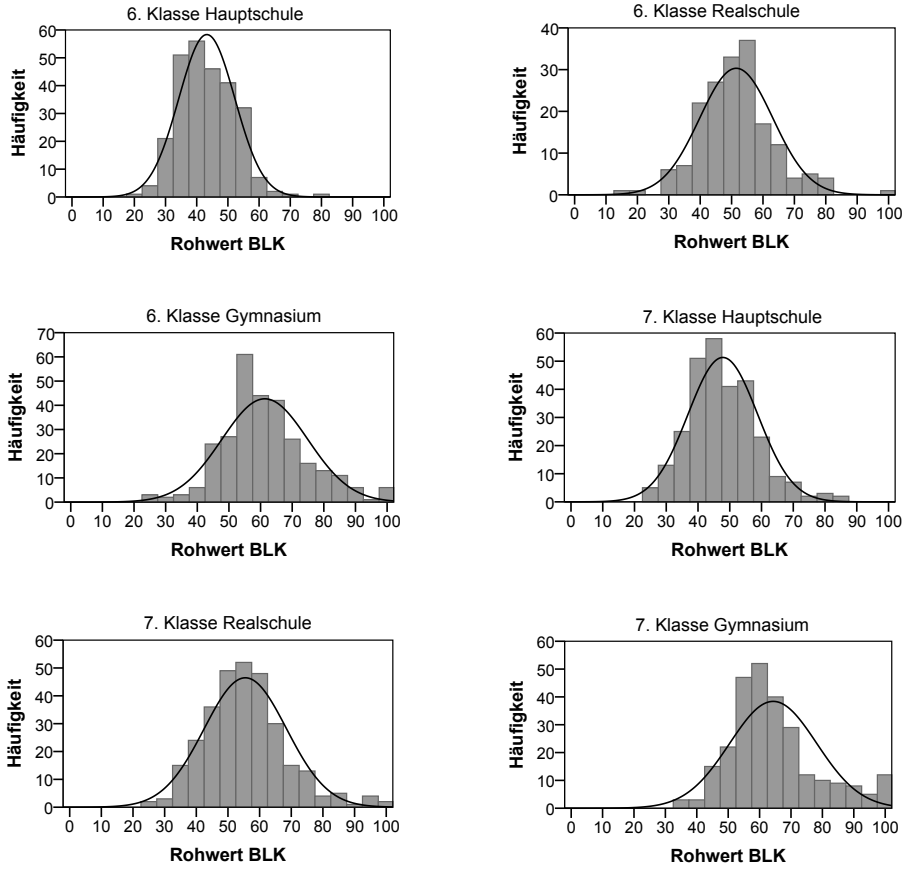


Abbildung 24. Rohwertverteilungen für den Subtest BLK von LESEN 6-7 auf Basis der Normdaten.

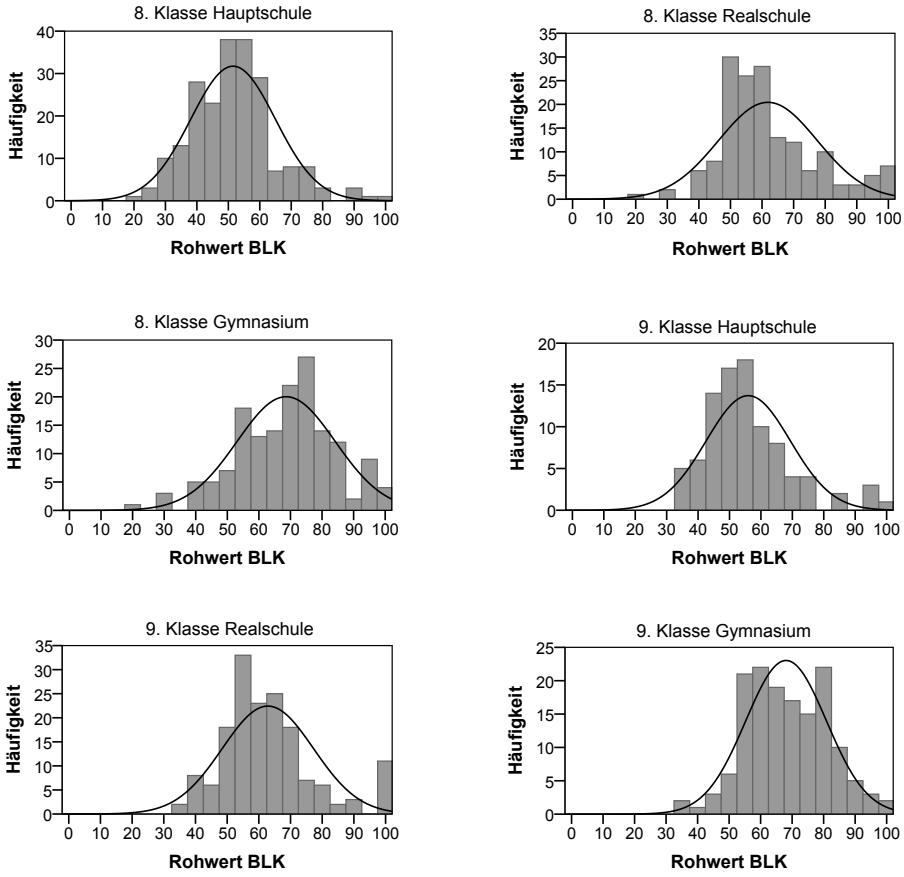


Abbildung 25. Rohwertverteilungen für den Subtest BLK von LESEN 8-9 auf Basis der Normdaten.

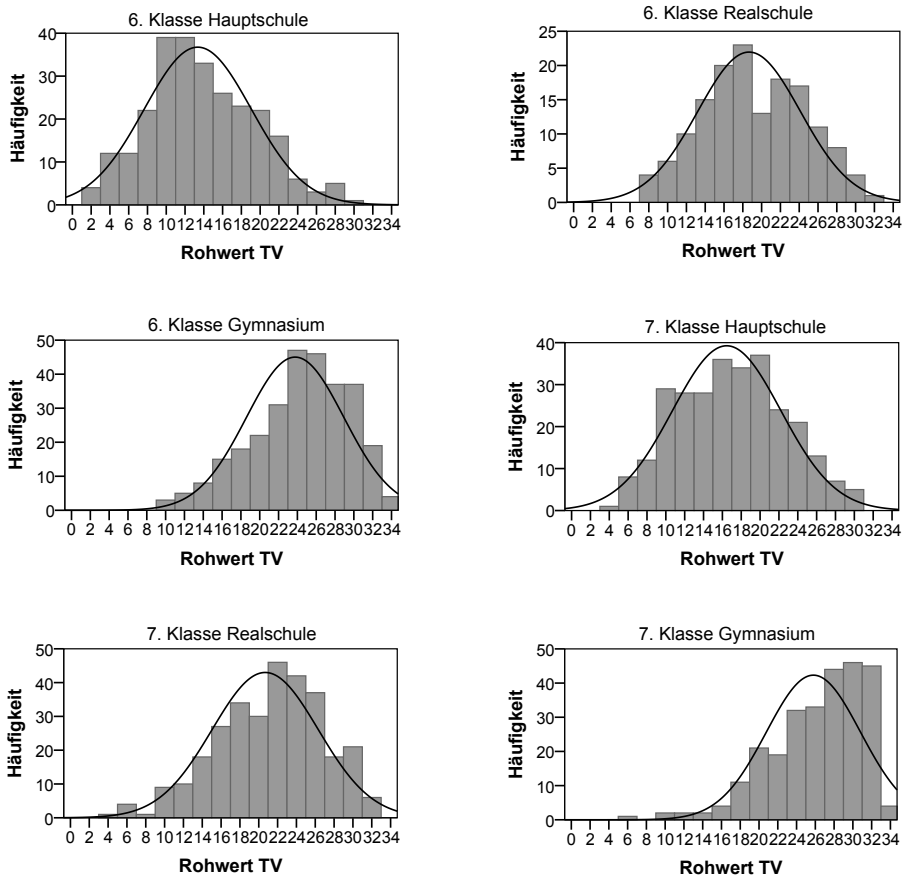


Abbildung 26. Rohwertverteilungen für den Subtest TV von LESEN 6-7 auf Basis der Normdaten.

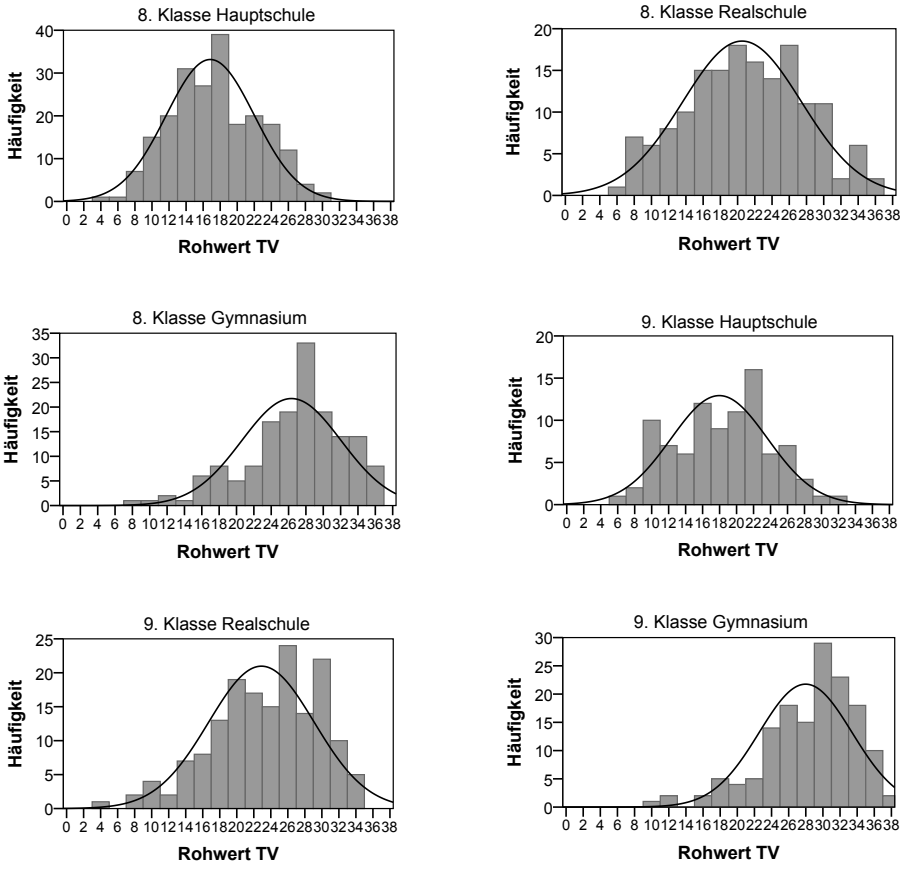


Abbildung 27. Rohwertverteilungen für den Subtest TV von LESEN 8-9 auf Basis der Normdaten.

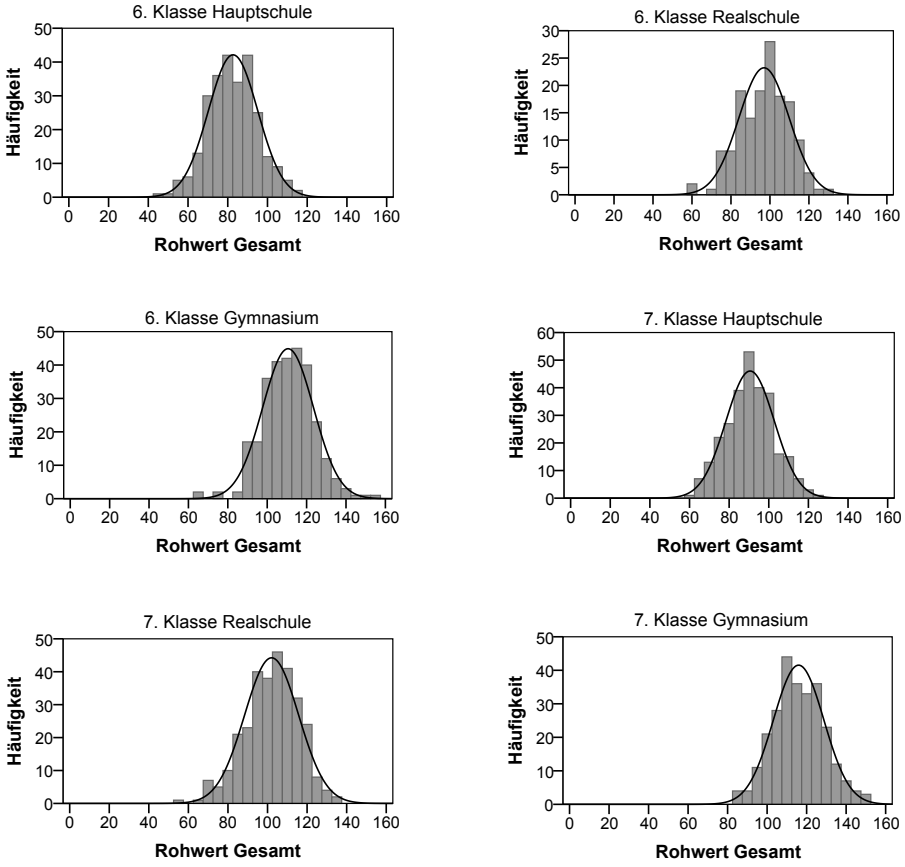


Abbildung 28. Rohwertverteilungen für das Gesamtergebnis von LESEN 6-7 auf Basis der Normdaten.

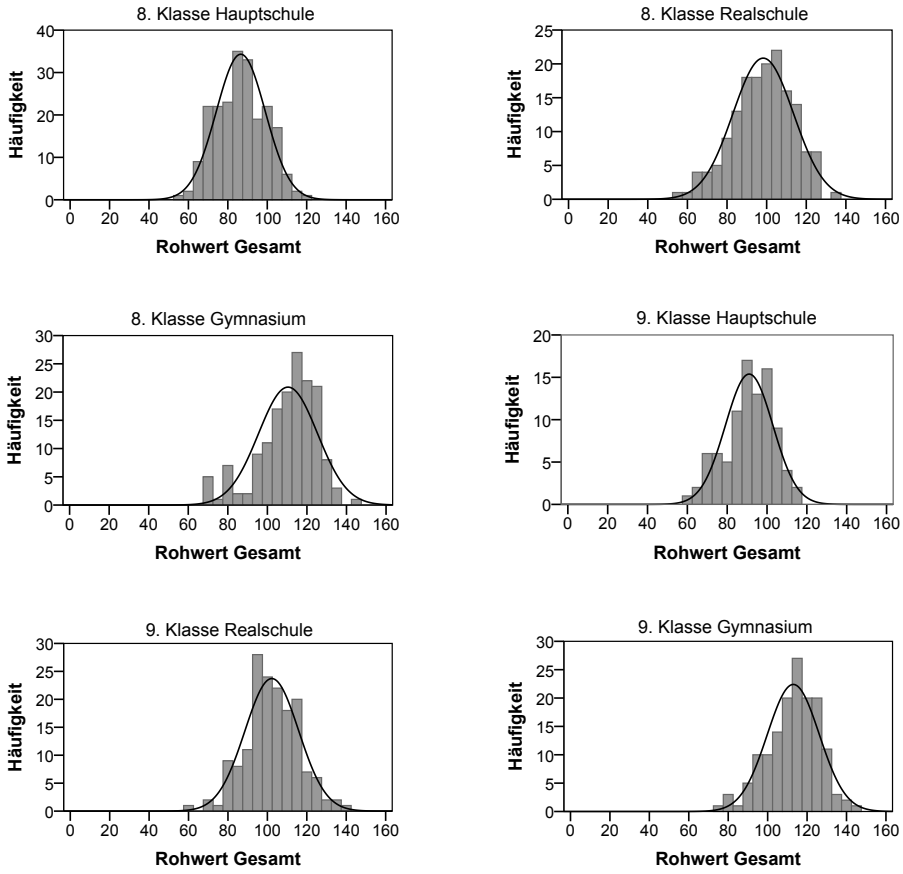


Abbildung 29. Rohwertverteilungen für das Gesamtergebnis von LESEN 8-9 auf Basis der Normdaten.

Anhang C: Itemanalysen und Modellpassung

Tabelle 42. Ergebnis der Itemanalyse zum Subtest BLK auf Basis der Normdaten der Klassenstufen sechs bis neun ($N = 2\,584$). In den Spalten aufgeführt sind jeweils Itemnummer (Item), Anzahl der Schüler, die das Item in Angriff nahmen ($N_{Angriff}$), und Anteil der Schüler, die das Item korrekt lösten in Prozent ($\%_{korr}$).

Item	$N_{Angriff}$	$\%_{korr}$	Item	$N_{Angriff}$	$\%_{korr}$	Item	$N_{Angriff}$	$\%_{korr}$
1	2 574	99.26	35	2 511	98.18	69	598	98.34
2	2 580	99.07	36	2 492	92.38	70	554	99.69
3	2 583	99.34	37	2 471	99.07	71	522	99.30
4	2 571	97.68	38	2 455	98.18	72	476	99.34
5	2 571	97.45	39	2 425	97.48	73	445	99.23
6	2 569	98.03	40	2 400	98.57	74	419	99.42
7	2 579	98.10	41	2 367	97.95	75	383	99.46
8	2 544	95.86	42	2 327	97.76	76	360	99.57
9	2 582	98.45	43	2 284	97.87	77	331	99.15
10	2 582	99.30	44	2 233	97.87	78	308	99.61
11	2 583	99.42	45	2 188	98.96	79	299	99.50
12	2 582	99.65	46	2 154	98.80	80	282	99.73
13	2 580	98.92	47	2 085	97.45	81	271	99.23
14	2 575	99.19	48	2 027	97.48	82	238	99.73
15	2 571	95.59	49	1 963	97.95	83	219	99.57
16	2 562	95.94	50	1 906	96.52	84	199	99.50
17	2 576	97.95	51	1 845	98.34	85	183	99.65
18	2 577	97.33	52	1 779	98.53	86	170	99.38
19	2 576	91.87	53	1 719	98.49	87	155	99.73
20	2 580	97.72	54	1 642	97.48	88	150	99.61
21	2 575	97.02	55	1 553	97.99	89	143	99.61
22	2 568	94.89	56	1 488	98.49	90	130	99.50
23	2 576	97.79	57	1 446	98.80	91	116	99.54
24	2 574	98.65	58	1 425	98.45	92	112	99.81
25	2 572	97.95	59	1 182	99.11	93	113	99.65
26	2 567	98.03	60	1 101	93.38	94	106	99.73
27	2 571	97.68	61	1 035	98.45	95	103	99.73
28	2 570	97.52	62	968	99.38	96	100	99.65
29	2 562	97.79	63	902	99.46	97	98	99.54
30	2 556	99.15	64	851	98.99	98	95	99.54
31	2 548	98.10	65	806	99.30	99	96	99.65
32	2 538	92.41	66	738	99.19	100	99	99.73
33	2 525	96.36	67	689	99.42	–	–	–
34	2 521	99.11	68	647	98.49	–	–	–

Tabelle 43. Ergebnis der Itemanalyse des Subtests TV auf Basis der Normdaten für LESEN 6-7 ($N = 1\,615$) und LESEN 8-9 ($N = 944$).

	LESEN 6-7				LESEN 8-9			
Item	p	SD	r_{it}	SK	p	SD	r_{it}	SK
1	.64	0.48	.36	0.38	.86	0.34	.28	0.40
2	.77	0.42	.42	0.50	.72	0.45	.42	0.47
3	.74	0.44	.42	0.48	.73	0.44	.36	0.41
4	.62	0.49	.28	0.29	.74	0.44	.46	0.52
5	.64	0.48	.40	0.42	.60	0.49	.34	0.34
6	.71	0.45	.40	0.44	.77	0.42	.25	0.30
7	.63	0.48	.47	0.48	.76	0.43	.29	0.34
8	.63	0.48	.46	0.47	.81	0.40	.28	0.36
9	.59	0.49	.41	0.42	.60	0.49	.27	0.28
10	.53	0.50	.38	0.38	.66	0.47	.19	0.20
11	.47	0.50	.34	0.34	.46	0.50	.36	0.36
12	.41	0.49	.39	0.40	.48	0.50	.37	0.37
13	.50	0.50	.40	0.40	.42	0.49	.42	0.43
14	.35	0.48	.40	0.42	.53	0.50	.38	0.38
15	.38	0.49	.34	0.35	.40	0.49	.32	0.33
16	.40	0.49	.42	0.43	.23	0.42	.17	0.20
17	.24	0.42	.23	0.27	.33	0.47	.23	0.24
18	.86	0.35	.35	0.50	.34	0.47	.33	0.35
19	.84	0.37	.32	0.43	.28	0.45	.29	0.32
20	.76	0.43	.27	0.31	.87	0.34	.37	0.54
21	.76	0.42	.42	0.49	.82	0.39	.34	0.44
22	.71	0.45	.40	0.44	.78	0.42	.50	0.60
23	.69	0.46	.42	0.45	.71	0.45	.37	0.41
24	.64	0.48	.44	0.46	.75	0.43	.45	0.53
25	.63	0.48	.38	0.39	.70	0.46	.45	0.49
26	.52	0.50	.27	0.27	.64	0.48	.41	0.43
27	.56	0.50	.28	0.28	.54	0.50	.43	0.43
28	.73	0.44	.42	0.48	.60	0.49	.47	0.48
29	.54	0.50	.35	0.36	.59	0.49	.43	0.43
30	.51	0.50	.24	0.24	.60	0.49	.45	0.46
31	.44	0.50	.43	0.43	.53	0.50	.50	0.50
32	.47	0.50	.35	0.35	.51	0.50	.38	0.38
33	.48	0.50	.45	0.45	.53	0.50	.31	0.31
34	.51	0.50	.40	0.40	.50	0.50	.36	0.36
35	–	–	–	–	.40	0.49	.29	0.30
36	–	–	–	–	.48	0.50	.25	0.25
37	–	–	–	–	.39	0.49	.30	0.30
38	–	–	–	–	.31	0.46	.29	0.31

Tabelle 44. Eigenwerte und Faktorladungen der Ein-Faktorenlösung für LESEN 6-7 und LESEN 8-9.

LESEN 6-7					LESEN 8-9				
Faktor	Eigenwert	Item	λ	h^2	Faktor	Eigenwert	Item	λ	h^2
1	10.13	1	0.50	0.25	1	10.48	1	0.46	0.21
2	1.66	2	0.61	0.37	2	1.84	2	0.59	0.34
3	1.19	3	0.60	0.36	3	1.48	3	0.51	0.26
4	1.12	4	0.39	0.15	4	1.31	4	0.65	0.43
5	1.06	5	0.55	0.30	5	1.23	5	0.46	0.22
6	1.04	6	0.56	0.32	6	1.20	6	0.37	0.14
7	1.00	7	0.63	0.40	7	1.10	7	0.42	0.17
8	0.99	8	0.62	0.38	8	1.07	8	0.43	0.19
9	0.96	9	0.56	0.31	9	1.05	9	0.37	0.14
10	0.94	10	0.51	0.26	10	1.00	10	0.25	0.06
11	0.88	11	0.46	0.21	11	0.95	11	0.49	0.24
12	0.87	12	0.54	0.29	12	0.92	12	0.50	0.25
13	0.81	13	0.55	0.30	13	0.90	13	0.57	0.32
14	0.77	14	0.56	0.31	14	0.87	14	0.52	0.27
15	0.77	15	0.48	0.23	15	0.84	15	0.45	0.20
16	0.73	16	0.57	0.33	16	0.79	16	0.26	0.07
17	0.72	17	0.34	0.12	17	0.77	17	0.32	0.10
18	0.67	18	0.58	0.34	18	0.74	18	0.46	0.21
19	0.64	19	0.51	0.26	19	0.73	19	0.42	0.18
20	0.64	20	0.39	0.15	20	0.70	20	0.62	0.38
21	0.62	21	0.61	0.37	21	0.68	21	0.54	0.29
22	0.58	22	0.56	0.31	22	0.65	22	0.74	0.55
23	0.56	23	0.58	0.33	23	0.63	23	0.54	0.29
24	0.54	24	0.61	0.38	24	0.60	24	0.66	0.44
25	0.53	25	0.52	0.27	25	0.57	25	0.64	0.41
26	0.51	26	0.37	0.13	26	0.53	26	0.59	0.34
27	0.48	27	0.39	0.15	27	0.49	27	0.57	0.33
28	0.45	28	0.60	0.36	28	0.48	28	0.65	0.42
29	0.43	29	0.48	0.23	29	0.47	29	0.58	0.34
30	0.40	30	0.32	0.10	30	0.43	30	0.61	0.37
31	0.38	31	0.59	0.35	31	0.41	31	0.68	0.46
32	0.35	32	0.47	0.22	32	0.39	32	0.51	0.26
33	0.33	33	0.61	0.37	33	0.37	33	0.41	0.17
34	0.29	34	0.54	0.29	34	0.36	34	0.49	0.24
–	–	–	–	–	35	0.30	35	0.41	0.17
–	–	–	–	–	36	0.28	36	0.35	0.12
–	–	–	–	–	37	0.23	37	0.41	0.17
–	–	–	–	–	38	0.19	38	0.40	0.16

Tabelle 45. Ergebnisse des auf Basis der Normdaten durchgeführten MAP-Tests.

LESEN 6-7			LESEN 6-7		
Eigenwert	Anzahl Komponenten	Vierte Potenz der mittleren Partialkorrelationen	Eigenwert	Anzahl Komponenten	Vierte Potenz der mittleren Partialkorrelationen
6.469	0	0.001	6.600	0	0.001
1.420	1	0.000	1.507	1	0.000
1.146	2	0.000	1.335	2	0.000
1.077	3	0.000	1.181	3	0.000
1.044	4	0.000	1.139	4	0.000
1.031	5	0.000	1.122	5	0.000
1.005	6	0.000	1.061	6	0.000
0.999	7	0.000	1.056	7	0.000
0.974	8	0.000	1.034	8	0.000
0.963	9	0.001	1.004	9	0.000
0.932	10	0.001	0.973	10	0.001
0.930	11	0.001	0.955	11	0.001
0.881	12	0.001	0.940	12	0.001
0.866	13	0.002	0.915	13	0.001
0.859	14	0.002	0.911	14	0.001
0.848	15	0.003	0.879	15	0.002
0.832	16	0.004	0.869	16	0.002
0.806	17	0.004	0.853	17	0.002
0.793	18	0.005	0.840	18	0.003
0.778	19	0.007	0.830	19	0.003
0.765	20	0.008	0.815	20	0.004
0.762	21	0.010	0.793	21	0.005
0.732	22	0.011	0.789	22	0.006
0.715	23	0.016	0.771	23	0.007
0.707	24	0.019	0.744	24	0.008
0.693	25	0.024	0.714	25	0.010
0.686	26	0.033	0.693	26	0.012
0.662	27	0.045	0.686	27	0.015
0.651	28	0.063	0.673	28	0.019
0.633	29	0.092	0.647	29	0.022
0.613	30	0.141	0.645	30	0.030
0.596	31	0.231	0.631	31	0.039
0.586	32	0.393	0.617	32	0.056
0.550	33	1.000	0.606	33	0.078
–	–	–	0.573	34	0.118
–	–	–	0.566	35	0.185
–	–	–	0.526	36	0.364
–	–	–	0.509	37	1.000

Tabelle 46. Item-Fit-Werte zur Beurteilung der Rasch-Modell-Konformität der Items von LESEN 6-7.

Item-Nr.	Item-parameter	Schätz-fehler	WMNSQ	WMNSQ-Konfidenzintervall Untergrenze	Obergrenze	*T*-Wert
1	-0.22	0.04	1.02	0.95	1.05	0.8
2	-0.98	0.04	0.94	0.94	1.06	-1.7
3	-0.82	0.04	0.95	0.94	1.06	-1.8
4	-0.12	0.04	1.09	0.96	1.04	4.0
5	-0.25	0.04	0.98	0.95	1.05	-0.9
6	-0.62	0.04	0.97	0.95	1.05	-1.0
7	-0.16	0.04	0.93	0.95	1.05	-3.3
8	-0.17	0.04	0.95	0.95	1.05	-2.4
9	-0.00	0.04	0.98	0.96	1.04	-1.0
10	0.33	0.04	1.01	0.96	1.04	0.7
11	0.61	0.04	1.05	0.96	1.04	2.2
12	0.89	0.04	1.00	0.96	1.04	0.0
13	0.46	0.04	0.99	0.96	1.04	-0.3
14	1.20	0.04	0.96	0.95	1.05	-1.7
15	1.03	0.04	1.04	0.96	1.04	1.7
16	0.95	0.04	0.96	0.96	1.04	-2.0
17	1.87	0.04	1.08	0.94	1.06	2.4
18	-1.69	0.05	0.95	0.90	1.10	-1.0
19	-1.53	0.04	0.99	0.91	1.09	-0.2
20	-0.90	0.04	1.07	0.94	1.06	2.3
21	-0.96	0.04	0.95	0.94	1.06	-1.6
22	-0.62	0.04	0.97	0.95	1.05	-1.0
23	-0.54	0.04	0.96	0.95	1.05	-1.4
24	-0.24	0.04	0.95	0.95	1.05	-2.2
25	-0.21	0.04	1.01	0.95	1.05	0.3
26	0.34	0.04	1.11	0.96	1.04	4.9
27	0.15	0.04	1.10	0.96	1.04	4.6
28	-0.77	0.04	0.94	0.94	1.06	-1.9
29	0.27	0.04	1.04	0.96	1.04	1.7
30	0.40	0.04	1.15	0.96	1.04	6.8
31	0.77	0.04	0.96	0.96	1.04	-1.9
32	0.58	0.04	1.03	0.96	1.04	1.4
33	0.57	0.04	0.94	0.96	1.04	-2.9
34	0.39	0.23	1.00	0.96	1.04	0.1

Tabelle 47. Item-Fit-Werte zur Beurteilung der Rasch-Modell-Konformität der Items von LESEN 8-9.

Item-Nr.	Item-parameter	Schätz-fehler	WMNSQ	WMNSQ-Konfidenzintervall Untergrenze	Obergrenze	*T*-Wert
1	-1.73	0.06	1.00	0.87	1.13	0.1
2	-0.68	0.06	0.95	0.93	1.07	-1.5
3	-0.77	0.06	1.00	0.93	1.07	-0.1
4	-0.81	0.06	0.91	0.92	1.08	-2.5
5	-0.04	0.05	1.03	0.95	1.05	1.2
6	-1.03	0.06	1.06	0.91	1.09	1.4
7	-0.92	0.06	1.04	0.92	1.08	1.1
8	-1.25	0.06	1.02	0.90	1.10	0.5
9	-0.07	0.05	1.09	0.94	1.06	3.3
10	-0.38	0.05	1.15	0.94	1.06	4.5
11	0.62	0.05	1.01	0.95	1.05	0.5
12	0.51	0.05	1.00	0.95	1.05	-0.1
13	0.81	0.05	0.95	0.95	1.05	-1.7
14	0.27	0.05	0.99	0.95	1.05	-0.2
15	0.89	0.05	1.04	0.95	1.05	1.4
16	1.82	0.06	1.12	0.92	1.08	2.8
17	1.24	0.05	1.10	0.94	1.06	3.2
18	1.22	0.05	1.01	0.94	1.06	0.5
19	1.55	0.06	1.05	0.93	1.07	1.3
20	-1.74	0.06	0.92	0.87	1.13	-1.2
21	-1.32	0.06	0.97	0.90	1.10	-0.6
22	-1.06	0.06	0.86	0.91	1.09	-3.3
23	-0.64	0.05	0.98	0.93	1.07	-0.5
24	-0.89	0.06	0.90	0.92	1.08	-2.6
25	-0.57	0.05	0.92	0.93	1.07	-2.4
26	-0.28	0.05	0.96	0.94	1.06	-1.3
27	0.25	0.05	0.96	0.95	1.05	-1.5
28	-0.04	0.05	0.91	0.95	1.05	-3.4
29	-0.02	0.05	0.95	0.95	1.05	-1.9
30	-0.07	0.05	0.94	0.94	1.06	-2.3
31	0.29	0.05	0.89	0.95	1.05	-4.2
32	0.36	0.05	1.00	0.95	1.05	0.2
33	0.28	0.05	1.07	0.95	1.05	2.6
34	0.42	0.05	1.01	0.95	1.05	0.5
35	0.88	0.05	1.07	0.95	1.05	2.5
36	0.54	0.05	1.11	0.95	1.05	4.1
37	0.95	0.05	1.06	0.94	1.06	2.0
38	1.39	0.33	1.03	0.93	1.07	1.0

Anhang D: Reliabilitätsanalysen

Tabelle 48. Reliabilitätswerte nach Klassenstufe sowie nach Klassenstufe und Schulart getrennt (N in Klammern).

		BLK		TV				GES
		KR-20	r_{tt}	KR-20	r_{tt}	EW	EAP/PV	r_{tt}
6	HS	.93 (263)	.68 (48)	.80 (263)	.86 (48)	.79 (263)	.80 (263)	.83 (48)
	RS	.95 (177)	.74 (42)	.77 (150)	.81 (42)	.75 (150)	.77 (150)	.86 (42)
	GYM	.96 (291)	.75 (48)	.78 (292)	.64 (48)	.75 (292)	.77 (292)	.82 (48)
	Gesamt	.96 (752)	.76 (138)	.87 (726)	.89 (138)	.86 (726)	.87 (726)	.90 (138)
7	HS	.95 (283)	.87 (35)	.79 (283)	.92 (14)	.78 (283)	.79 (283)	.93 (14)
	RS	.96 (304)	.89 (32)	.79 (304)	.78 (32)	.77 (304)	.80 (304)	.77 (32)
	GYM	.97 (267)	.86 (51)	.78 (265)	.89 (51)	.71 (265)	.76 (265)	.86 (51)
	Gesamt	.97 (891)	.88 (140)	.86 (889)	.85 (119)	.83 (889)	.85 (889)	.87 (119)
8	HS	.96 (215)	.92 (35)	.72 (215)	.78 (35)	.70 (215)	.72 (215)	.76 (35)
	RS	.97 (162)	.85 (36)	.87 (162)	.88 (36)	.86 (162)	.88 (162)	.90 (36)
	GYM	.97 (156)	.89 (44)	.79 (155)	.88 (44)	.76 (155)	.79 (155)	.85 (44)
	Gesamt	.97 (539)	.89 (139)	.86 (538)	.88 (139)	.85 (539)	.85 (539)	.88 (139)
9	HS	.96 (92)	.74 (17)	.76 (92)	.67 (17)	.75 (92)	.76 (92)	.39 (17)
	RS	.97 (163)	.89 (12)	.81 (163)	.92 (12)	.80 (163)	.82 (163)	.93 (12)
	GYM	.95 (148)	.80 (41)	.81 (148)	.87 (16)	.76 (148)	.80 (148)	.86 (16)
	Gesamt	.97 (406)	.86 (83)	.86 (406)	.87 (58)	.84 (406)	.86 (406)	.84 (58)

Anmerkungen:
KR-10: Interne Konsistenz auf Basis der Kuder-Richardson-20-Formel
r_{tt}: Retest-Reliabilitätswerte mit Korrektur für die Varianzeinschränkung
EW: Erwartungswertmethode
EAP/PV: expected a posteriori/plausible value

Anhang E: Validitätsanalysen

Tabelle 49. Häufigkeiten der einzelnen Rohwerte in der Normstichprobe beim Subtest BLK.

Rohwert	Häufigkeit	Rohwert	Häufigkeit	Rohwert	Häufigkeit
17	1	47	58	74	43
19	2	48	63	75	19
21	2	49	78	76	17
23	3	50	57	77	17
24	5	51	68	78	20
25	2	52	63	79	23
26	2	53	79	80	22
27	3	54	78	81	23
28	5	55	71	82	6
29	12	56	86	83	15
30	10	57	100	84	14
31	5	58	112	85	12
32	15	59	90	86	9
33	16	60	59	87	11
34	22	61	53	88	7
35	24	62	57	89	13
36	20	63	56	90	8
37	34	64	42	91	4
38	28	65	63	92	4
39	34	66	43	93	5
40	37	67	41	94	5
41	39	68	41	95	3
42	44	69	41	96	2
43	60	70	33	97	13
44	50	71	38	98	9
45	47	72	32	99	18
46	68	73	22	100	28

Tabelle 50. Fehlerhäufigkeit der Normstichprobe im Subtest BLK.

Fehlerzahl	Häufigkeit	Fehlerzahl	Häufigkeit
0	1.031	18	2
1	742	19	3
2	379	20	1
3	166	21	1
4	99	23	3
5	46	24	2
6	18	25	3
7	17	26	1
8	15	27	1
9	7	29	2
10	10	30	1
11	6	32	2
12	9	33	1
13	2	35	1
14	3	39	1
15	4	41	1
16	1	42	1
17	2	–	–

Tabelle 51. Anzahl nicht bearbeiteter Items in der Normstichprobe im Subtest BLK.

Anzahl nicht in Angriff genommener Items	Häufigkeit	Anzahl nicht in Angriff genommener Items	Häufigkeit	Anzahl nicht in Angriff genommener Items	Häufigkeit
0	68	26	38	52	65
1	9	27	27	53	56
2	4	28	31	54	62
3	7	29	44	55	41
4	3	30	32	56	42
5	5	31	49	57	44
6	7	32	47	58	50
7	3	33	41	59	37
8	1	34	44	60	35
9	2	35	67	61	32
10	14	36	48	62	27
11	16	37	52	63	16
12	4	38	61	64	24
13	12	39	67	65	16
14	12	40	77	66	12
15	11	41	62	67	14
16	20	42	230	68	11
17	16	43	44	69	8
18	18	44	48	70	3
19	30	45	67	71	10
20	15	46	86	72	3
21	22	47	64	73	2
22	15	48	68	75	4
23	18	49	64	76	2
24	26	50	71	77	2
25	21	51	57	80	1

Tabelle 52. Ergebnisse der KS-Tests auf Normalverteilung (*KS-Z*).

Klasse	BLK	TV	GES	LGVT-LG	LGVT-LV	MKT	WLST
6	1.74*	1.15*	0.66	2.35*	1.12*	1.27*	–
7	0.95	1.09*	0.52	1.00	0.97	–	1.43*
6-7	1.52*	1.46*	–	2.34*	1.59*	–	–
8	1.03	1.16*	0.67	1.42*	0.74	–	1.34*
9	0.64	1.17*	0.89	1.29*	0.78	–	1.00
8-9	0.96	1.64*	–	1.69*	0.95	–	1.87*

Anmerkung: *: $p < .20$, keine Normalverteilung angenommen.

Tabelle 53. Deskriptive Statistik für die Analysen zu den Hypothesen H3 bis H5 für LESEN 6-7.

Klassen-stufe	Schul-art	Ge-schlecht	*N*	Subtest BLK		Subtest TV		GES	
				M	*SD*	*M*	*SD*	*M*	*SD*
6	HS	m	120	42.34	9.64	13.95	5.64	82.69	13.02
		w	135	44.10	8.37	12.98	5.76	82.92	11.98
		Gesamt	255	43.27	9.01	13.44	5.71	82.81	12.45
	RS	m	82	50.17	11.41	18.45	5.59	95.15	12.69
		w	63	54.97	12.49	18.71	5.44	99.17	13.06
		Gesamt	145	52.26	12.09	18.57	5.51	96.89	12.96
	GYM	m	144	60.89	14.10	24.24	4.96	110.99	13.24
		w	128	62.29	13.21	23.46	5.44	110.83	12.83
		Gesamt	272	61.55	13.68	23.87	5.20	110.91	13.02
	Gesamt	m	346	51.92	14.53	19.30	6.99	97.42	17.96
		w	326	53.34	13.94	18.20	7.29	97.02	17.74
		Gesamt	672	52.61	14.26	18.77	7.15	97.22	17.84
7	HS	m	156	45.50	10.83	16.30	6.04	88.47	12.26
		w	119	50.55	10.66	16.82	5.22	93.40	11.81
		Gesamt	275	47.69	11.03	16.53	5.70	90.61	12.29
	RS	m	156	53.25	13.29	20.87	5.77	100.63	13.76
		w	143	57.99	12.41	20.52	5.58	103.67	13.61
		Gesamt	299	55.52	13.07	20.70	5.67	102.08	13.75
	GYM	m	131	62.69	14.39	25.74	4.92	114.71	12.66
		w	132	65.99	13.30	25.86	5.16	117.26	12.82
		Gesamt	263	64.35	13.92	25.80	5.03	115.99	12.78
	Gesamt	m	443	53.31	14.55	20.70	6.78	100.51	16.65
		w	394	58.42	13.67	21.19	6.45	105.13	15.98
		Gesamt	837	55.72	14.36	20.93	6.63	102.68	16.49
Gesamt	HS	m	276	44.13	10.43	15.28	5.97	—	—
		w	254	47.13	10.03	14.78	5.83	—	—
		Gesamt	530	45.56	10.34	15.04	5.91	—	—
	RS	m	238	52.19	12.74	20.03	5.81	—	—
		w	206	57.06	12.48	19.97	5.59	—	—
		Gesamt	444	54.45	12.84	20.00	5.70	—	—
	GYM	m	275	61.75	14.24	24.95	4.99	—	—
		w	260	64.17	13.36	24.68	5.42	—	—
		Gesamt	535	62.93	13.86	24.82	5.20	—	—
	Gesamt	m	789	52.70	14.55	20.08	6.90	—	—
		w	720	56.12	14.01	19.84	7.00	—	—
		Gesamt	1509	54.33	14.39	19.97	6.95	—	—

Tabelle 54. Deskriptive Statistik für die Analysen zu den Hypothesen H3 bis H5 für LESEN 8-9.

Klassen-stufe	Schul-art	Ge-schlecht	*N*	Subtest BLK		Subtest TV		GES	
				M	*SD*	*M*	*SD*	*M*	*SD*
8	HS	m	112	47.21	11.54	17.03	5.35	83.79	11.64
		w	96	56.17	14.05	16.80	4.97	89.83	12.65
		Gesamt	208	51.34	13.50	16.92	5.17	86.58	12.46
	RS	m	96	60.90	15.69	20.54	7.01	97.83	15.22
		w	57	64.09	15.15	20.44	6.97	99.45	15.80
		Gesamt	153	62.08	15.52	20.50	6.97	98.43	15.41
	GYM	m	93	65.02	15.70	25.90	5.85	107.54	15.96
		w	55	74.40	13.63	27.25	5.71	115.47	12.20
		Gesamt	148	68.51	15.60	26.41	5.82	110.49	15.13
	Gesamt	m	301	57.08	16.23	20.89	7.07	95.61	17.31
		w	208	63.16	16.05	20.56	7.17	99.25	17.06
		Gesamt	509	59.56	16.42	20.76	7.11	97.09	17.28
9	HS	m	35	54.91	12.84	20.74	5.84	93.88	11.23
		w	48	56.67	12.68	16.46	5.00	89.61	11.65
		Gesamt	83	55.93	12.70	18.27	5.74	91.41	11.60
	RS	m	85	62.78	15.69	22.29	6.77	101.31	14.72
		w	72	63.04	13.10	23.79	5.45	103.71	12.35
		Gesamt	157	62.90	14.51	22.98	6.23	102.41	13.70
	GYM	m	59	63.63	13.13	27.49	6.40	109.82	14.83
		w	89	71.03	11.78	28.30	4.68	115.13	11.57
		Gesamt	148	68.08	12.82	27.98	5.43	113.01	13.18
	Gesamt	m	179	61.52	14.65	23.70	7.00	102.66	15.20
		w	209	64.98	13.66	24.03	6.79	105.33	15.44
		Gesamt	388	63.38	14.21	23.88	6.88	104.10	15.37
Gesamt	HS	m	147	49.04	12.27	17.91	5.67	—	—
		w	144	56.33	13.57	16.69	4.96	—	—
		Gesamt	291	52.65	13.41	17.31	5.36	—	—
	RS	m	181	61.78	15.68	21.36	6.94	—	—
		w	129	63.50	13.99	22.31	6.37	—	—
		Gesamt	310	62.50	15.00	21.76	6.71	—	—
	GYM	m	152	64.48	14.72	26.52	6.10	—	—
		w	144	72.32	12.58	27.90	5.10	—	—
		Gesamt	296	68.29	14.25	27.19	5.67	—	—
	Gesamt	m	480	58.73	15.79	21.94	7.17	—	—
		w	417	64.07	14.91	22.30	7.19	—	—
		Gesamt	897	61.22	15.61	22.11	7.18	—	—

Tabelle 55. Ergebnisse der KS-Tests für alle Zellen der ANOVA zur Prüfung der Normalverteilungsannahme bezüglich des Gesamtergebnisses von LESEN 6-7 und LESEN 8-9 für H3 bis H5.

Klassenstufe	Schulart	Geschlecht	*N*	*KS-Z*
6	HS	m	120	0.36
		w	135	0.50
	RS	m	82	0.64
		w	63	0.43
	GYM	m	144	0.61
		w	128	0.79
7	HS	m	156	0.60
		w	119	0.74
	RS	m	156	0.62
		w	143	0.69
	GYM	m	131	0.44
		w	132	0.45
8	HS	m	112	0.59
		w	96	0.48
	RS	m	96	0.55
		w	57	0.60
	GYM	m	93	1.06
		w	55	0.60
9	HS	m	35	0.77
		w	48	0.63
	RS	m	85	0.41
		w	72	0.63
	GYM	m	59	0.85
		w	89	0.83

Anmerkung: alle $p > .20$

Tabelle 56. Deskriptive Statistik für die Analysen zur Hypothese H6 für LESEN 6-7 und LESEN 8-9.

Klassen-			Subtest BLK		Subtest TV		GES	
stufe	LRS	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
6	Ja	82	44.84	11.99	16.60	6.72	88.29	16.02
	Nein	632	53.73	13.96	19.18	6.99	98.65	17.15
7	Ja	79	48.14	13.69	18.70	7.45	93.62	17.61
	Nein	792	56.25	14.06	20.98	6.53	103.20	15.94
8	Ja	44	49.14	12.53	17.91	7.12	86.21	14.74
	Nein	482	60.83	16.30	21.07	6.97	98.34	16.93
9	Ja	25	57.04	13.77	18.56	6.93	92.69	14.26
	Nein	377	63.68	14.20	24.09	6.70	104.60	15.04

Tabelle 57. Deskriptiver Vergleich der Schüler mit Deutsch als Muttersprache vs. anderer Muttersprache hinsichtlich Beruf und Schulabschluss der Eltern.

LESEN 6-7				
Deutsch	Beruf	180	2.65	0.95
Andere	Beruf	17	2.24	1.20
Deutsch	Schule	180	2.16	0.76
Andere	Schule	17	2.12	0.93
LESEN 8-9				
Muttersprache		N	M	SD
Deutsch	Beruf	194	2.55	0.78
Andere	Beruf	23	1.96	0.37
Deutsch	Schule	191	2.29	0.65
Andere	Schule	20	1.75	0.91

Tabelle 58. Korrelationen möglicher Kontrollvariablen mit den Subtests und dem Gesamtergebnis von LESEN 6-7 und LESEN 8-9 für die Prüfung von H7 (Daten der Validierungsstichprobe).

	LESEN 6-7				LESEN 8-9			
	BLK	TV	GES 6. Kl.	GES 7. Kl.	BLK	TV	GES 8. Kl.	GES 9. Kl.
N	197	197	92	105	192	192	110	82
Klassenstufe	.14*	.33**	—	—	.31**	.33**	—	—
Schulart	.67**	.43**	.60**	.78**	.56**	.51**	.67**	.45**
Geschlecht	.13	.05	.03	.14	.29**	.11	.07	.18
Beruf Eltern	.26**	.23**	.37**	.33**	.06	.15*	.09	.09
Schule Eltern	.32**	.24**	.34**	.44**	.27**	.30**	.32**	.29**
LRS	-.23**	-.01	-.07	-.20*	.18*	.11	.16	.22*

Anmerkung: *: $p < .05$; **: $p < .01$

Mit LESEN 6-7 und LESEN 8-9 wurden zwei Lesetests für die Sekundarstufe entwickelt, die auf dem aktuellen Forschungsstand basieren und sowohl die basale Lesekompetenz als auch das Textverständnis überprüfen. Die Testkonstruktion berücksichtigte neben der Klassischen Testtheorie auch die Item Response Theorie. Beide Tests wurden an einer umfangreichen Stichprobe normiert und auf Reliabilität und Validität sowie hinsichtlich weiterer Testgütekriterien überprüft. LESEN 6-7 und LESEN 8-9 erfüllen die gängigen Gütekriterien in sehr zufriedenstellendem Maße. Sie ermöglichen nicht nur auf Gruppen-, sondern auch auf Individualebene eine umfassende und zuverlässige Erfassung des Leseverständnisses von Sekundarschülern und differenzieren in allen vier Klassenstufen im gesamten Leistungsspektrum.

ISBN 978-3-95826-008-5
9 783958 260085

Würzburg University Press