technische universität dortmund

fakultät rehabilitationswissenschaften

ENTWICKLUNG UND ERFORSCHUNG INKLUSIVER BILDUNGSPROZESSE

Masterthesis

Grenzen des Messens bei der Lernverlaufsdiagnostik im Lesen

Eine Analyse der Boden- und Deckeneffekte in den Jahrgängen 3/4 mit der Online-Plattform Levumi

vorgelegt von

Denise König

denise.koenig@tu-dortmund.de

MA Lehramt für sonderpädagogische Förderung (LABG 2009)

Betreuende: **Prof. Dr. Markus Gebhardt**

Prof. Dr. Jörg-Tobias Kuhn

eingereicht am: **23.07.2018**

I Inhaltsverzeichnis

II Zusammenfassung

Infolge des Inkrafttretens des Artikels 24 der UN-Behindertenrechtskonvention im Jahr 2009 steht das deutsche Bildungssystem vor großen Herausforderungen. Insbesondere durch die zunehmende Heterogenität, entstehend durch den gemeinsamen Unterricht von Schülerinnen und Schülern mit und ohne Förderbedarf, nimmt die Relevanz von Unterrichtsprinzipien wie die Differenzierung und die Individualisierung stetig zu. Aufgrund dessen steigt auch die Relevanz der Lernverlaufsdiagnostik, da diese auf eine optimale Anpassung zwischen Lernausgangslage der Schülerschaft und Fördermaßnahmen abzielt, immer weiter an. Ein Beispiel für ein Verfahren der Lernverlaufsdiagnostik ist das Diagnoseinstrument Levumi. Die Online-Lernplattform stellt ein neues Diagnoseinventar dar und umfasst Tests für mehrere Unterrichtsfächer. Die vorliegende Arbeit fokussiert sich ausschließlich auf die Domäne der Lesetests innerhalb der Levumi-Plattform und untersucht, inwiefern die Ergebnisse der Tests durch Decken- und Bodeneffekte beeinflusst werden. Dabei werden die Lesetestergebnisse der Jahrgangsstufen 3 und 4 einer inklusiven Grundschule in NRW betrachtet. In aller Kürze kann bezüglich der Ergebnisse festgehalten werden, dass in den Lesetests tatsächlich Boden- und Deckeneffekte auftreten; das Ausmaß dieser wird aber als nicht gravierend eingeschätzt. Insbesondere ist der Test zum sinnentnehmenden Lesen von diesen betroffen, während der Test zum Silben lesen frei von Boden- und Deckeneffekten ist.

1 Einleitung

> Bisher lag der diagnostische Schwerpunkt – vor dem Hintergrund des bestehenden Schulsystems – in erster Linie darin, Förderbedarfe zu fixieren und ggf. Ressourcen für die (sonder-) pädagogische Förderung an die Schülerin bzw. den Schüler zu binden. In den Hintergrund rückten Ableitungen von gezielten Fördermaßnahmen sowie deren Evaluation. Zukunftsweisend und praxisrelevant erscheint daher der Ansatz der lernbegleitenden Diagnostik (Voß & Gebhardt 2017, S. 95).

Dieses Zitat legt nahe, dass besonders im Hinblick auf ein inklusives Schulsystem ein Wandel von der traditionellen Statusdiagnostik hin zu individuellen, lernbegleitenden Diagnostikformen bereits stattfindet bzw. in Zukunft verstärkt stattfinden wird. Denn die Lernverlaufsdiagnostik bietet die Möglichkeit, Fördermaßnahmen individueller an den Lernstand der Schülerinnen und Schüler anzupassen und erhält dadurch die besondere Relevanz im inklusiven Bildungssystem und folglich auch in der Bildungsforschung.

1.1 Problemstellung

Die Lernverlaufsdiagnostik stellt im deutschsprachigen Raum einen relativ neuen Forschungsansatz dar. Dementsprechend mangelt es bisher an Diagnoseinstrumenten, die den Lernverlauf valide abbilden können (vgl. Diehl 2010, S. 74). Ein erst in den letzten Jahren entwickeltes Verfahren zur Lernverlaufsdiagnostik ist die Online-Plattform Levumi. Levumi ist ein offenes Forschungsprojekt, welches in engem Austausch zwischen Administratoren und Lehrkräften stetig weiterentwickelt wird und auch bestehende Tests werden stets überarbeitet. Da „die Forschung zu Onlinetestungen und Lernverlaufsmessung noch in den Kinderschuhen steckt“ (Gebhardt, Diehl & Mühling 2016b, S. 450), sind einige Fragen bezüglich der praktischen Anwendung von Levumi noch unbeantwortet. Aufgrund dessen beschäftigt sich auch die vorliegende Arbeit mit einer dieser noch offenen Fragestellungen. Denn diese Arbeit betrachtet das Diagnoseinstrument Levumi hinsichtlich der Grenzen des Messens in der Lernverlaufsdiagnostik. Das bedeutet, dass die ermittelten Daten auf das Auftreten von Boden- und Deckeneffekten hin untersucht werden, da ein ausgeprägtes Auftreten dieser dafür sorgt, dass Leistungen in den extremen Merkmalsbereichen nicht differenziert gemessen werden können (vgl. Moosbrugger & Kelava 2012, S. 138).

1.2 Zielsetzung

Mithilfe einer durchgeführten Langzeitstudie in den Jahrgangsstufen 3/4 einer inklusiven Grundschule in NRW soll aufgedeckt werden, in welchem Umfang die durch die Online-Plattform Levumi ermittelten Lernverläufe der Schülerinnen und Schüler durch Boden- und Deckeneffekte beeinflusst werden, d.h. wann die Grenzen des Messens erreicht werden. Die

Langzeitstudie erstreckt sich über vier Messzeitpunkte innerhalb eines gesamten Schuljahres. Innerhalb dieser Studie wird sich auf die Lerndomäne des Lesens beschränkt. Demnach werden Tests zum Unterrichtsfach Mathematik außen vorgelassen. So werden drei Leseflüssigkeitstests (Silben-, Wörter-, Pseudowörterlesen) und ein Test zum sinnentnehmenden Lesen auf Boden- und Deckeneffekte hin überprüft.

1.3 Vorgehensweise

In der folgenden Arbeit wird zunächst das im Hinblick auf die empirische Untersuchung bedeutsame theoretische Hintergrundwissen dargelegt. Dazu werden beginnend in Kapitel 2 relevante Elemente der Lesekompetenz aufgeführt. Dies umfasst beispielsweise die Unterscheidung zwischen Lesetechnik und sinnerfassendem Lesen sowie die Vorstellung von relevanten Leseentwicklungsmodellen.

Daraufhin folgt im dritten Kapitel der Hauptteil des theoretischen Hintergrundes, d.h. grundlegende Kenntnisse zur Lernverlaufsdiagnostik werden dargelegt. Im Besonderen wird hier die Relevanz der Lernverlaufsdiagnostik für die inklusive Schulpraxis sowie die Bedeutung der Boden- und Deckeneffekte in den Lernverläufen thematisiert. Da sich auf Lernverlaufstests zur Lesefähigkeit beschränkt wird, wird außerdem ein Überblick über den aktuellen Forschungsstand der Lernverlaufsdiagnostik im Lesen dargelegt.

Nachdem die Fragestellung dieser Arbeit offengelegt wurde (vgl. Kapitel 4), wird versucht diese mithilfe der empirischen Untersuchung zu beantworten. Der empirische Teil beginnt mit der Operationalisierung der Hypothesen bzw. der Operationalisierung der Fragestellung. Daran anschließend wird das Forschungsdesign - die Lernplattform Levumi - sowie die Untersuchungsstichprobe, das Untersuchungssetting, die Durchführung und das Vorgehen der empirischen Auswertung umfassend beschrieben (vgl. Kapitel 5).

Anschließend werden in Kapitel 6 und 7 die Untersuchungsergebnisse interpretiert, strukturiert und diskutiert. Das letzte und achte Kapitel liefert abschließend eine kurze Zusammenfassung sowie einen Ausblick auf einen möglichen weiteren Umgang mit den erhaltenen Ergebnissen.

THEORETISCHER TEIL

2 Entwicklung der Lesekompetenz

> Die guten Leutchen wissen nicht, was es einem für Zeit und Mühe gekostet, um *lesen zu lernen*. Ich habe achtzig Jahre dazu gebraucht und kann noch jetzt nicht sagen, daß ich am Ziele wäre (Johann Wolfgang von Goethe 1830).[1]

Gemäß diesem Goethe-Zitat wird in dem folgenden Kapitel die, Goethe zur Folge, sehr komplexe Entwicklung der Lesekompetenz dargestellt. Nach einer kurzen Einführung zur Situation der Leseförderung an deutschen Schulen und zu dem Stellenwert des Lesens in unserer Gesellschaft, zielt dieses Kapitel darauf ab, zu klären, was überhaupt unter der allgemeinen Lesekompetenz zu verstehen ist, wie es sich mit unterschiedlichen Begriffen zur Lesefertigkeit und zur Lesefähigkeit verhält und welche Vorläufer- und Teilfähigkeiten zur Entwicklung dieser vorhanden sein sollten. Des Weiteren wird die bereits angesprochene komplexe Entwicklung der Lesekompetenz betrachtet. Dabei wird vorzugsweise die Leseentwicklung in der dritten und vierten Jahrgangsstufe in den Blick genommen, da diese Altersklasse auch für die durchgeführte Langzeitstudie ausgewählt wurde. In Anbetracht des in der empirischen Studie verwendeten Forschungsdesigns, d.h. die Online-Plattform Levumi, wird der Fokus auf Theorien zur Leseentwicklung gesetzt, welche auch in diesem Diagnoseinstrument berücksichtigt wurden.

Dass die Lesefähigkeit deutscher Schülerinnen und Schüler[2] im Vergleich zu anderen Ländern unterdurchschnittlich ausgebildet ist, wurde als Folge der „alarmierenden Ergebnisse der internationalen Leistungsvergleichsstudien des letzten Jahrzehnts“ und insbesondere durch den „PISA-Schock“ des Jahres 2001 bewiesen (Garbe 2010, S. 9). Ergebnisse der ersten PISA-Studie im Jahr 2000 zeigten, dass „beinahe jeder vierte getestete deutsche Schüler […] enorme Schwierigkeiten beim Lesen (und somit auch beim Textverstehen)“ aufzeigte (Hochstadt, Krafft & Olsen 2015, S. 115). Ursächlich dafür ist unter anderem, dass der Leseförderung lange Zeit keine hohe Relevanz zugeschrieben wurde, da man die Annahme verfolgte, dass wichtige Teilkompetenzen, wie beispielsweise die Leseflüssigkeit, am Ende der Grundschulzeit bereits vollständig ausgebildet seien (vgl. ebd., S. 115). Auch in der PISA-Studie von 2009 blieben die Lesefähigkeiten deutscher SuS trotz Verbesserungen weiterhin unter dem OECD-Durchschnitt (vgl. Wilckens 2018, S. 147). Inzwischen hat sich der Terminus der Lesekompetenz daher zu einem Schlüsselbegriff der Fachdidaktik Deutsch, aber

[1] Eckermann, Johann Peter (1836). Gespräche mit Goethe in den letzten Jahren seines Lebens (Band 2). Leipzig: Brockhaus.

[2] Zugunsten einer besseren Lesbarkeit und einer verkürzten Schreibweise wird in dieser Arbeit anstelle des Ausdrucks „Schülerinnen und Schüler“ die Abkürzung „SuS“ verwendet, da auf diese Weise beide Geschlechter genannt werden und dennoch eine verkürzte Schreibweise möglich ist.

auch des gesamten Bildungssystems entwickelt (vgl. Spinner 2010, S. 48). Denn die Lesekompetenz hat nicht nur einen immensen Einfluss auf die Schulleistungen in allen Unterrichtsfächern, sondern bestimmt auch über den gesamten Bildungserfolg der SuS (vgl. Schneider 2017, S. 73). Das Lesen stellt somit „eine der Schlüsselqualifikationen unserer Gesellschaft dar“, da es nicht nur für die Alltagsbewältigung, sondern auch für das Berufsleben unerlässlich ist (Frey 2010, S. 15). Ebenso spricht der Lehrplan NRW für den Deutschunterricht an Grundschulen von einer „Schlüsselfunktion“ des Lesens (MSW 2008, S. 26). Folglich trägt die Beherrschung der Lesefähigkeit zur Teilhabe am gesellschaftlichen Leben bei und gilt daher auch als eine Kulturtechnik (vgl. Günthner 2013, S. 11).
Aufgrund der immer noch zu verbessernden Lesefähigkeiten deutscher SuS und der unumstrittenen Bedeutung dieser, sollte an deutschen Schulen weiterhin eine umfangreiche und verbesserte Leseförderung etabliert werden. Eine Möglichkeit dies zu erreichen ist die Lernverlaufsdiagnostik, die im dritten Kapitel vorgestellt wird.

2.1 Begriffsklärungen

Eine in der Fachliteratur häufig angeführte Definition des Lesens stammt von Friedrich Kainz (1956). Er versteht unter dem Lesen „das verstehende Aufnehmen von schriftlich fixierten Sprachfügungen, somit die auf Grund der erworbenen Kenntnis der Schriftzeichen vollzogene Tätigkeit des Sinnerfassens graphisch niedergelegter Gedankengänge“ (ebd., S. 162). Diese Definition beinhaltet sowohl die elementaren Prozesse wie das Erkennen von Graphemen und Worten, als auch die Sinnentnahme des Gelesenen (vgl. Scheerer-Neumann 2006a, S. 513). Außen vor lässt Kainz jedoch die Übertragung des Gelesenen in die Lautsprache, was ebenfalls einen wichtigen Teilprozess des Lesens darstellt.

Der Begriff des Lesens beschreibt folglich einen hochkomplexen Prozess, der unterschiedliche Leistungen ganz unterschiedlicher Gehirnareale erfordert (vgl. Bertschi-Kaufmann 2011, S. 8). So werden für den Leseprozess beispielsweise Gehirnareale beansprucht, die „visuelle/auditive Wahrnehmungsleistungen sowie mentale Verarbeitungs- und Wiedergabeprozesse“ zeitgleich bewältigen müssen (Günthner 2013, S. 35). Für geübte Leser[3] stellt der Leseprozess kaum noch eine Herausforderung dar; Wörter werden als Ganzes erkannt, da diese bereits im Sichtwortschatz abgespeichert sind. Diesen Sichtwortschatz haben Leseanfänger jedoch noch nicht aufgebaut, sodass sie die meisten Wörter erst mühevoll erlesen müssen, da ihre Lesekompetenz noch nicht automatisiert ist (vgl. Marx 2007, S. 18f.). Die

[3] Zugunsten einer besseren Lesbarkeit wird an dieser Stelle sowie im weiteren Verlauf dieser Arbeit auf die doppelte Nennung von Personenbezeichnungen, mit Ausnahme der bereits genannten Abkürzung „SuS“, zur Kennzeichnung beider Geschlechtsformen verzichtet. Jedoch werden stets alle Geschlechtsformen mitbedacht.

Komplexität des Leseprozesses lässt erahnen, dass der Erwerb der Lesekompetenz ebenfalls eine hochkomplexe Aufgabe darstellt.

In der Schriftspracherwerbsforschung wird im Hinblick auf den Leseerwerb zwischen den Termini der Lesefähigkeit, der Lesefertigkeit und der Lesekompetenz unterschieden. Die Begriffe der Lesefähigkeit und der Lesekompetenz werden in der Literatur uneinheitlich verwendet, daher wird sich in dieser Arbeit auf folgende Begriffsbedeutungen geeinigt:

Die *allgemeine Lesekompetenz* sowie die *Lesefähigkeit* werden in den folgenden Ausführungen synonym verwendet. Diese Begriffe unterscheiden sich aber von dem der *Lesefertigkeit*. Denn die Lesefertigkeit bezeichnet die Lesetechnik und ist somit als eine Komponente der Lesekompetenz zu verstehen. Die zweite Komponente der Lesekompetenz bildet das Leseverständnis. Schneider (2017) erklärt somit folgerichtig, „dass mit der Lesefertigkeit und dem Leseverständnis insgesamt zwei grundlegende Komponenten der Lesekompetenz zu unterscheiden sind, die in der Entwicklung aufeinanderfolgen und deutliche Verbindungen aufweisen“ (S. 22). Auch das Modell „simple view of reading“ von Gough & Tunmer (1986) stimmen mit dieser Auffassung überein. So geht das Modell von folgender Aussage aus: „Reading equals the product of decoding and comprehension“ (S. 7). Demzufolge entwickelt sich das Leseverständnis („reading“) aus dem Zusammenspiel der Dekodierfähigkeit („decoding“) und dem allgemeinen Sprachverständnis („comprehension“) (vgl. Wember 2012, S. 194).

Gough & Tunmer entsprechend benennt Wember (2012) die *Lesefertigkeit* auch als die „Dekodierfähigkeit“ (S. 194). Demnach beschreibt die Lesefertigkeit die Fähigkeit, Grapheme in Phoneme umsetzen zu können. Diese wird daher als der technische Akt innerhalb des Leseprozesses angesehen (vgl. Dehn 2010, S. 145). Scheerer-Neumann führt an, dass die Teilkomponente der Lesefertigkeit beherrscht wird, wenn „der Leser beim leisen Lesen von Fließtexten eine Lesegeschwindigkeit von ungefähr 100 bis 150 Wörtern pro Minute erreicht und […] automatisiert lesen kann“ (Scheerer-Neumann 2015, S, 73).

Garbe zufolge kommt der Leseflüssigkeit innerhalb der Lesefertigkeit darüber hinaus eine besondere Bedeutung zu. Sie beschreibt diese eben nicht nur als eine Teilkompetenz der Lesefertigkeit, sondern in Anlehnung an die angelsächsische Forschung als „bridge between decoding and comprehension“ (Garbe 2010, S. 18). Demnach hat die Leseflüssigkeit eine Doppelfunktion: einerseits ermöglicht sie die Automatisierung der Dekodierprozesse, andererseits unterstützt sie die Automatisierung der Verstehensprozesse (vgl. Holle 2009, S. 147). Denn erst, wenn das Arbeitsgedächtnis durch ein automatisiertes und flüssiges Dekodieren von Graphemen zu Phonemen entlastet wird, ist das Arbeitsgedächtnis bereit den Sinn des Gelesenen zu konstruieren (vgl. Garbe 2010, S. 18). Zu beachten ist aber, dass auch unter

der Leseflüssigkeit verschiedene Dimensionen gefasst werden. Während beispielsweise Dehn (2010) die Leseflüssigkeit als eine Teilkompetenz der Lesefertigkeit, neben der Lesegeschwindigkeit und der -genauigkeit ansieht, versteht Sturm (2011) unter der Leseflüssigkeit einen übergeordneten Begriff, der vier Dimensionen (das genaue Dekodieren, das automatisierte Dekodieren, Lesetempo, Leseausdruck) umfasst. Nach Sturm besteht der Zusammenhang zwischen Leseflüssigkeit und Leseverständnis außerdem lediglich in der Grundschule und nimmt mit steigendem Alter ab (vgl. S. 15). Das *Leseverständnis* umfasst die komplexere Teilfähigkeit der Lesekompetenz. Die Sinnkonstruktion oder Sinnentnahme beschreibt die Fähigkeit den inhaltlichen Sinn aus dem Text erfassen zu können (vgl. Frey 2016, S. 16). Aufgrund der Komplexität gilt das Leseverständnis auch als „high-order-“, die Leseflüssigkeit als „low-order-Prozess“ (vgl. Holle 2009, S. 107).

Bezüglich der Entwicklung beider Teilbereiche ist die Forschung sich einig, dass sie sich während der Grundschulzeit nicht parallel entwickeln, sich aber wechselseitig bedingen (vgl. Dehn 2010, S.146). Bei Betrachtung des Lehrplans Deutsch für die Grundschulen in NRW wird schnell offensichtlich, dass das Leseverstehen im Mittelpunkt steht (vgl. MSW 2008, ab S. 31). Die Lesefertigkeit gerät dort hingegen in den Hintergrund, obwohl diese als Grundlage für das Verständnis angesehen wird und nicht als bereits gegeben angesehen werden sollte, sondern ebenfalls der Förderung bedarf.

Während Bildungsstandards so häufig einen Lesebegriff im weiteren Sinne verfolgen, da Interpretationen und das aktive Verarbeiten des Gelesenen im Fokus stehen (vgl. Schenk 2007, S. 12) und sie neben dem Verständnis von literarischen Texten auch die Anwendung von Lesestrategien oder das Entwickeln eigener Texte in die Lesefähigkeit miteinbeziehen (vgl. Spinner 2010, S. 52), konzentriert sich diese Arbeit im Hinblick auf das spätere Erhebungsinstrument auf ein Verständnis der Lesefähigkeit im engeren Sinne. Das bedeutet, es werden die grundlegenden Prozesse der Lesefertigkeit und der Sinnentnahme, wie beispielsweise das Erkennen und Dekodieren der Grapheme und Worte sowie die reine Sinnentnahme ohne weitere Verarbeitung dessen fokussiert (vgl. Schenk 2007, S. 12).

2.1.1 Vorläuferfähigkeiten und Teilkompetenzen

Um die oben beschriebene Lesekompetenz auszubilden, sind bestimmte Vorläuferfähigkeiten, d.h., Fähigkeiten, die sich positiv auf die Entwicklung der Lesefähigkeit auswirken, von grundsätzlicher Bedeutung. Die Ausprägung von Vorläuferfähigkeiten kann bereits Aufschluss über die folgende Schulleistungsentwicklung geben. Vorläuferfähigkeiten gelten somit auch als Ansatzpunkt für Präventionsmaßnahmen, um bereits frühzeitig das Auftreten von Schriftspracherwerbsproblemen zu vermeiden (vgl. Ennemoser et al. 2012, S. 53).

Schon bei Schuleintritt unterscheidet sich die Ausprägung der Lesekompetenz von den SuS untereinander sehr stark (vgl. Schneider 2017, S. 17). Aufgrund von verschiedenen Bedingungen und Einflussfaktoren während der Kindheit, können Vorläuferfähigkeiten, wie der Wortschatz, die visuelle Aufmerksamkeit oder das Arbeitsgedächtnis, ganz verschieden ausgeprägt sein (vgl. Jungjohann et al. 2017, S. 6).

Vorläuferfähigkeiten können in spezifische und unspezifische Vorläuferfähigkeiten geteilt werden. Unspezifische Vorläuferfähigkeiten umfassen Fähigkeiten, die nicht konkret mit dem Leseerwerb zusammenhängen, diesen aber positiv bedingen. Dazu gehören Komponenten wie die Konzentrationsfähigkeit, Intelligenz, Lernfreude, das Selbstkonzept oder die Motivation (vgl. Marx 2007, S. 38f.). Eine besondere Stellung unter den unspezifischen Vorläuferfähigkeiten nimmt das Interesse ein. Denn schon früh machen die SuS, unabhängig von der schulischen Förderung, wichtige Erfahrungen mit der Schrift (vgl. Scheerer-Neumann 2006a, S. 513). Diese Erfahrungen ermöglichen die Ausbildung eines Interesses an der Schriftsprache und ihrer Funktion. Durch Beobachtungen und Imitationen lesender Personen in ihrem Umfeld erhalten die Kinder Einsicht in elementare Regularitäten des Schriftsystems, wie z.B. die Schreibrichtung oder das Zeilenprinzip, welche als grundlegende Vorläuferfähigkeiten angesehen werden können (vgl. Braun 2010, S. 175).

Dem gegenüber stehen die spezifischen Fertigkeiten, die direkt mit dem Leseprozess zusammenhängen. Dazu gehören beispielsweise visuelle Wahrnehmungsleistungen. Leser müssen über eine ausreichend ausgebildete Figur-Grund-Wahrnehmung und Formkonstanz verfügen. Außerdem sollten sie die visuelle Formdiskrimination beherrschen, damit sie ähnlich erscheinende Grapheme wie <b, d, p> unterscheiden können (vgl. ebd., S. 177). Für den Verstehensprozess müssen die SuS aktiv auf bereits vorhandenes Wissen, wie z.B. auf vorher gelesene Sätze oder Textpassagen, zurückgreifen können und dieses Wissen gleichzeitig weiter ausbauen können (vgl. Bredel, Fuhrhop & Noack 2011, S. 74).

Allgemein versteht man unter der phonologischen Bewusstheit die Fähigkeit, dass der Fokus weg von dem Inhalt hin zu den lautlichen Aspekten der Sprache gesetzt werden kann (vgl. Forster 2005, S. 37). Die phonologische Bewusstheit wird häufig als elementare spezifische Vorläuferfähigkeit bezeichnet, ist aber auch eine ebenso kontrovers diskutierte Vorläuferfähigkeit. Denn die Forschung ist sich derzeit noch uneinig, ob die phonologische Bewusstheit tatsächlich als eine Vorläuferfähigkeit des Schriftspracherwerbs anzusehen ist, oder ob sie als Konsequenz aus der Leseentwicklung entsteht (vgl. Schneider 2017, S. 57). Sie wird unterschieden in die phonologische Bewusstheit im engeren und im weiteren Sinne. Phonologische Bewusstheit im weiteren Sinne beschäftigt sich mit größeren sprachlichen Einheiten,

d.h. mit Wörtern, Reimen oder Silben. Es kann davon ausgegangen werden, dass diese Fähigkeiten bereits vor dem Schuleintritt erworben werden. Im Gegensatz dazu steht die phonologische Bewusstheit im engeren Sinne. Diese umfasst beispielsweise die Lautdiskrimination innerhalb eines Wortes (vgl. Marx 2007, S. 45). In Bezug auf die fachdidaktische Diskussion kann durch die Unterscheidung zwischen engerem und weiterem Sinne der Kompromiss getroffen werden, dass sich die Fertigkeiten des engeren Bewusstheitsbegriffs erst durch die gezielte schulische Förderung entwickeln, während der weitere Bewusstheitsbegriff als Vorläuferfähigkeit angesehen werden kann.

Eine Studie von Ennemoser et al. (2012) konnte die wichtigsten Prädikatoren für die Entwicklung der Lesekompetenz identifizieren. Hinsichtlich der Lesegeschwindigkeit konnten sie beweisen, dass die Benenngeschwindigkeit, d.h. der Zugriff auf das semantische Lexikon, besonders zu Beginn der Grundschulzeit als das stärkste Vorhersagemerkmal angesehen werden kann. Aber auch die phonologische Bewusstheit spielt zu Beginn der Grundschulzeit eine große Rolle, diese nimmt dann jedoch stetig ab. Ähnliche Befunde wurden auch im Hinblick auf das Satz- und Textverständnis herausgestellt. Hier muss jedoch auch die hohe Relevanz der linguistischen Kompetenz betont werden, die besonders im Hinblick auf das Textverständnis gegen Ende der Grundschulzeit weiter zunimmt. Als wichtigste Erkenntnis wird aus dieser Studie somit gezogen, dass die linguistische Kompetenz, d.h. der Wortschatz, stärker berücksichtigt werden sollte, da diese ähnlich wie die phonologische Bewusstheit einen großen Einfluss auf die Lesekompetenzentwicklung ausübt (vgl. S. 61f.).

2.2 Entwicklung der Lesekompetenz

Wie bereits zu Beginn des Kapitels 2.1.1 angesprochen wurde, steigen die SuS mit sehr heterogenen Vorerfahrungen bezüglich der Lesekompetenz in die Schule ein. Die weitere schulische Entwicklung der Lesekompetenz baut sich dann auf Basis der bereits gemachten Erfahrungen und der bisherigen Lesesozialisation in der Familie weiter aus, sodass auch die weitere Entwicklung folglich von einer großen Heterogenität geprägt ist (vgl. Garbe 2010, S. 16).

Den vielfältigen Ausgangslagen der Lesekompetenz entsprechend verfolgt also auch die weitere Entwicklung keine lineare Abfolge, sondern verläuft eher sprunghaft, sodass Leseentwicklungsmodelle, als idealtypische Darstellungen, lediglich als eine Orientierungsmöglichkeit dienen können, um den aktuellen Lernstand eines Kindes besser einschätzen zu können, nicht aber als Beschreibung der tatsächlichen Entwicklungsschritte aller SuS (vgl. Bertschi-Kaufmann 2010, S. 28).

Der Lehrplan Deutsch für Grundschulen in NRW beschreibt die Leseentwicklung folgendermaßen: „Buchstabenverbindungen [werden] geläufig und auch simultan erkannt. Die ursprüngliche Sinnerwartung wird im Prozess der Texterschließung bestätigt, modifiziert, überprüft und in Beziehung zum Vorwissen gesetzt“ (MSW 2008, S. 26). Als das Ziel der Leseentwicklung kann folglich die automatisierte Lesekompetenz bezeichnet werden. Denn durch ein automatisiertes Lesen, im Sinne von einem direkten oder auch simultanen Worterkennen anstelle des lautorientierten Erlesens, nimmt die Gehirntätigkeit sowie die Belastung des Arbeitsgedächtnisses ab, sodass mehr Raum für die eigentliche Sinnkonstruktion besteht (vgl. Dehn 2010, S. 141). Die Unterscheidung zwischen dem direkten Worterkennen und dem lautorientierten Erlesen eines Wortes wird auch im „Dual-Route-Model“ von Coltheart (1978) beschrieben. Denn Coltheart stellt in seinem Modell zwei unterschiedliche Wege des Worterkennens auf und entwickelte somit eine bedeutsame Theorie zum Wortlesen (vgl. Gebhardt, Diehl & Mühling 2016a, S. 6). Obwohl das Modell mehrfach als zu stark vereinfachend kritisiert wurde, stellt es die zwei grundsätzlichen Möglichkeiten des Worterkennens, d.h. den zentralen Prozess des Leseerwerbs, adäquat dar (vgl. Marx 2007, S 19).

Coltheart unterscheidet zwischen dem lautorientierten (indirekten) Lesen einerseits, welches der alphabetischen Strategie entspricht und dem direkten Worterkennen andererseits (vgl. Scheerer-Neumann 2006a, S. 515). Wird ein Wort mithilfe des indirekten Verarbeitungsmechanismus erlesen, ist das Wort noch nicht im Sichtwortschatz präsent. Dies ist besonders am Anfang des Leseerwerbs der Fall. Daher erlesen SuS anfangs Buchstabe für Buchstabe und synthetisieren die in Phoneme übertragenen Grapheme erst im Anschluss zu einem gesamten Wort (vgl. Wilckens 2018, S. 148). Um sequenziell Graphem für Graphem zu erlesen und in Laute übertragen zu können, müssen die SuS über die visuelle Diskriminationsfähigkeit sowie über die Phonem-Graphem-Korrespondenzen verfügen (vgl. Marx 2007, S. 19.) Teilweise können durch das lautorientierte Worterlesen sogenannte Wortvorformen entstehen, die von der gewohnten Aussprache, aufgrund der extremen Dehnung einzelner Laute, abweichen. SuS erkennen das tatsächliche Wort daher erst durch die Suche nach adäquat klingenden Wörtern. Wurde in ihrem inneren Lexikon ein ähnliches Wort gefunden, beginnt der inhaltliche Verstehensprozess und die SuS können das Wort auch artikulatorisch korrekt verbalisieren. Stehen die zu erlesenen Worte nicht isoliert, sondern im Rahmen eines Textes oder einer Geschichte, stellt der Kontext eine gute Hilfe zum schnellen Worterkennen dar (vgl. Scheerer-Neumann 2006a, S. 516). Eine weitere Eigenschaft des Leseanfängers ist das „halblaute Lesen“. Leseanfänger lesen, selbst wenn sie für sich lesen, verbal vor, während der geübte Leser die Prozesse komplett verinnerlichen kann. Der Anfänger benötigt den Wortklang jedoch noch als Unterstützung zur Sinnfindung (vgl. Schenk 2007, S. 21). Der

indirekte synthetische Leseweg ermöglicht das Lesen unbekannter Wörter. In Bezug auf die im Diagnoseinstrument Levumi verwendeten Pseudowörter ist der indirekte Weg von fundamentaler Bedeutung, da diese Wörter entsprechend der Phonem-Graphem-Korrespondenzen erlesen werden müssen und nicht als Ganzes erkannt werden können (vgl. Marx 2007, S. 19). Folglich lässt sich der aktuelle Lernstand des synthetischen Verarbeitungsmechanismus durch das Vorlesen von Pseudowörtern und der Lernstand des lexikalischen Verarbeitungsmechanismus durch das Vorlesen bekannter Wörter ermitteln (vgl. Wilckens 2018, S. 149). Der direkte oder lexikalische Weg des Worterkennens beruht auf der Wiedererkennung auffälliger visueller Eigenschaften der Buchstabenabfolgen. Hier ist keine Synthese der Phoneme notwendig, da das Wort bereits als Ganzes aufgenommen wird. Dieser Weg ist nur möglich, wenn das Wort im orthographischen Lexikon bereits gespeichert ist. Im Gegensatz zu Colthearts damaligen Annahmen steht nun die Überzeugung dafür, dass beide Verarbeitungswege beim Erlesen eines Wortes gleichzeitig tätig werden können (vgl. Gebhardt et al., 2016a, S. 6f.).

Im Hinblick auf die Untersuchungsstichprobe werden in den weiteren Ausführungen besonders die Jahrgangsstufen 3 und 4 fokussiert, obwohl die Leseförderung innerhalb der gesamten ersten vier Schuljahre sehr entscheidend für die Entwicklung der allgemeinen Lesekompetenz ist (vgl. Garbe 2010, S. 17).

Während zu Beginn des Anfangsunterrichts hauptsächlich das nichtlexikalische Lesen von Bedeutung ist, nimmt die Entwicklung der lexikalischen Lesefähigkeit stetig zu. Die SuS nehmen mit der Zeit immer größere Einheiten, wie Silben, Morpheme und schließlich ganze Wörter, als eins wahr und werden folgend direkt als eins erkannt. Dadurch steigen während der gesamten Grundschulzeit die Leseflüssigkeit sowie die Lesegeschwindigkeit stetig an. Die Zunahme der Leseflüssigkeit ist dabei von besonderer Bedeutung, da sie die weitere Leseentwicklung stark beeinflusst (vgl. Jungjohann, Gegenfurtner & Gebhardt 2018a, S. 102). Die Spracherwerbsforschung geht davon aus, dass die Lesekompetenz beherrscht wird, sobald sich das Leseverständnis dem Hörverständnis zeitlich gesehen annähert. Bereits während des zweiten Schuljahres lässt sich eine Verringerung der Diskrepanzen zwischen Lese- und Hörverständnis wahrnehmen, da sich die Lesefertigkeiten stetig verbessern. Eine annähernde Angleichung der beiden Verständnisprozesse findet jedoch erst während des vierten Schuljahres statt (vgl. Scheerer-Neumann 2006a, S. 523).

Nach Garbe (2010) ist es von elementarer Bedeutung, neben der Förderung des nichtlexikalischen und des lexikalischen Lesens, die Lesemotivation der SuS von Anfang an nicht zu vernachlässigen, sondern stets mit zu fördern (vgl. S. 16). Demzufolge ist es nennenswert,

dass, den Untersuchungen zur Lesemotivation in der Grundschule von Richter und Plath (2012) zufolge, im Hinblick auf die Lesemotivation besonders das dritte Schuljahr ein sehr wesentliche Entwicklungszeit darstellt. Denn bereits in der dritten Klasse ist ein erster Abfall der Lesemotivation bezüglich Büchern und Geschichten zu verzeichnen (vgl. S. 43). Des Weiteren belegt die genannte Studie, dass der bedeutsamste Faktor für die Einstellung gegenüber dem Lesen in der Familie liegt, dass die Schule durch den Deutschunterricht aber ebenso Einfluss nehmen kann. Jedoch nehmen das Interesse und die Freude am Deutschunterricht wiederum ab dem dritten Schuljahr stetig ab. Nach Richter und Plath (2012) kann daher die „Klassenstufe 3 [als] eine Art Umschlagpunkt" bezeichnet werden, da es sich hier bereits entschieden hat, ob die SuS nach dem Erwerb der basalen Lesekompetenz „den persönlichen Wert des Lesens tatsächlich erfahren" haben (vgl. S. 75).

2.2.1 Frith: The Six-step Model of skills in Reading and Writing Acquisition

Die Entwicklung der Lesekompetenz wird häufig in Leseentwicklungsmodellen dargestellt. Obwohl sie, wie zu Beginn des Kapitels 2.2 bereits erläutert, lediglich idealtypische Beschreibungen abgeben, unterstützen sie den aktuellen Lernstand eines Kindes zu ermitteln und sind dadurch auch für diese Ausführungen relevant. Denn obwohl Entwicklungsmodelle die Entwicklungsprozesse stark vereinfachen, geben sie dennoch relevante Gesetzmäßigkeiten des Erwerbsprozesses wieder (vgl. Diehl 2011, S. 168).

Ein für die Forschung grundlegendes Stufenmodell, auf dem eine Vielzahl weiterer Modelle aufbaut, ist das „Six-step Model of skills in Reading and Writing Acquisition" bzw. das „Sechs-Stufen-Modell des Erwerbs von Lese- und Schreibfertigkeiten" von Uta Frith (1985), welches aufgrund der Relevanz in der Forschung im Folgenden kurz dargestellt wird (vgl. Marx 2007, S. 26). Abbildung 1 verdeutlicht den Zusammenhang zwischen dem rezeptiven Leseprozess und dem produktiven Schreibprozess. In dieser Arbeit wird lediglich die Leseentwicklung fokussiert. Daher bleiben Friths Aussagen zur Entwicklung des Schreibens im Weiteren unbeachtet und es wird nur die linke Spalte von Abbildung 1 in den Blick genommen.

Step	*Reading*	*Writing*
1a	*logographic*$_1$	(symbolic)
1b	logographic$_2$ →	logographic$_2$
2a	logographic$_3$	*alphabetic*$_1$
2b	alphabetic$_2$ ←	alphabetic$_2$
3a	*orthographic*$_1$	alphabetic$_3$
3b	orthographic$_2$ →	orthographic$_2$

Abbildung 1: *„Six-step Model of skills in Reading and Writing Acquisition" (entnommen aus Frith 1985, S. 311)*

Wie in Abbildung 1 deutlich wird, teilt Frith den Leseentwicklungsprozess in drei Phasen ein, die jeweils in mehreren Niveaustufen auftauchen, sodass insgesamt sechs Schritte entstehen (vgl. Frith 1985, S. 310). Diese drei Phasen lassen sich auch in den meisten späteren Entwicklungsmodellen in ähnlicher Form wiederfinden (vgl. Diehl 2011, S. 167). Frith geht nicht davon aus, dass alle SuS die Stufen in der gleichen Zeit durchlaufen oder dass alle Phasen die gleiche Zeit benötigen, sodass das Durchlaufen der Stufenabfolge zeitweise auch rückläufig sein kann (vgl. Gehle-Davids 2015, S. 11). Vor der logographischen Stufe befinden sich die Kinder in der „pre-literacy phase“, welche nicht in das Stufenmodell mit aufgenommen wurde, in dieser Phase beschäftigen sich die Kinder bereits mit metalinguistischen Begriffen wie „*Wörter*“ oder „*Sätze*“ (vgl. ebd., S. 308). In der logographischen Stufe angekommen erkennen die Kinder bereits vertraute Wörter (z.B. den eigenen Namen, <MAMA, PAPA, OMA>) oder Logos. Auf dieser Stufe erlesen sie die Wörter noch nicht buchstabenweise, sondern erkennen die Wortbedeutung aufgrund optischer Auffälligkeiten (vgl. Schründer-Lenzen 2009, S. 30). In der daran anschließenden alphabetischen Strategie erlernen die SuS die Phonem-Graphem-Korrespondenzen. Voraussetzungen dafür sind folglich die Beherrschung der Buchstaben sowie Einsicht in die phonologische Struktur unserer Sprache. Da nun Worte Buchstabe für Buchstabe bzw. Laut für Laut zusammengeschliffen und synthetisiert werden, sind die SuS ab dieser Phase auch in der Lage Pseudowörter zu erlesen (vgl. Marx 2007, S. 28). Der dritte und bei Frith letzte Schritt ist die Verwendung der orthographischen Strategie. Diese beinhaltet das automatisierte Erlesen von Wörtern durch orthographische Einheiten, beispielsweise durch Morpheme oder Silben. Ab dieser Phase wird nicht mehr jedem einzelnen Buchstaben ein Laut zugeordnet, sondern ganze Buchstabenkombinationen werden als eins erkannt (vgl. Scheerer-Neumann 2006a, S. 517).

Eine der Erweiterungen des Modells von Frith stammt von Klaus B. Günther (1995). Er geht davon aus, dass ausreichend belegt ist, dass Friths Modell die tatsächlich ablaufenden Entwicklungsprozesse adäquat wiedergibt. Dennoch präzisiert und erweitert er das Modell Friths zu einem fünf-phasigen Entwicklungsmodell (vgl. S. 98f.). Im Unterschied zu Frith nimmt Günther die präliteral-symbolische Strategie mit in sein Modell auf (s. Abbildung 2). Diese umfasst in Anlehnung an Friths „preliteracy phase“ die Vorbedingungen des Schriftspracherwerbs, wie beispielsweise das Bildlesen oder Nachahmungen von Lese- und Schreibprozessen (vgl. ebd., S. 100f.).

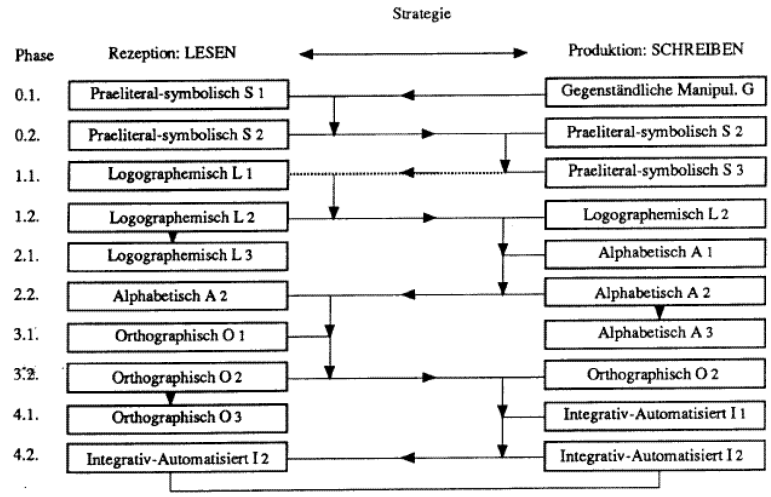


Abbildung 2: *Modell der Aneignung der schriftlichen Sprache als mehrphasiger, strategiebestimmter Entwicklungsprozess (entnommen aus Günther 1995, S. 99)*

Entscheidend für diese Zwecke ist jedoch die integrativ automatisierte Phase, die Günther (1995) als die letzte Phase des Erwerbsprozesses ansieht. Diese umfasst zwar keinen neuen Entwicklungsschritt, verdeutlicht aber, dass die bisherigen Schritte gefestigt und automatisiert werden müssen und dass der Leseerwerb noch nicht mit dem Durchlaufen der orthographischen Phase abgeschlossen ist (vgl. S. 109). SuS erreichen diese Stufe, sobald die Leseprozesse soweit gefestigt sind, dass sie ohne Stocken oder langes Nachdenken erfolgen können (vgl. Gehle-Davids 2015, S. 16).

2.2.2 Klicpera et al.: Kompetenzentwicklungsmodell des Lesens

Im Unterschied zu dem Modell Friths oder Günthers gibt das Kompetenzentwicklungsmodell des Lesens von Klicpera, Schabmann und Gasteiger-Klicpera (2013) vordergründig keine Abfolge bestimmter Erwerbsphasen vor, sondern vielmehr einen Überblick über Kompetenzen, die im Leseprozess erworben werden müssen. Wie in Abbildung 3 erkenntlich wird, unterscheidet dieses Modell, wie das „Dual-Route-Model“ von Coltheart (1978), zwischen nichtlexikalischen und lexikalischen Lesezugängen. Beide Zugangsweisen entwickeln sich durch die Instruktion, d.h. durch den schulischen Unterricht oder durch andere Fördermaßnahmen (vgl. Klicpera, Schabmann und Gasteiger-Klicpera 2013, S. 31). Betont wird, dass diese Teilkompetenzen nicht nacheinander aufgebaut werden, sondern sich als parallele Lesestrategien ausbilden (vgl. Jungjohann 2018a, S. 102).

Des Weiteren unterscheiden sie zwischen der präalphabetischen und der alphabetischen Phase. Die präalphabetische Stufe, die Vorstufe des Modells, entspricht in ihren Grundzügen der logographischen bzw. logographemischen Stufe nach Frith und Günther. Worte werden nicht aufgrund der Buchstabensequenzen erlesen, sondern aufgrund von visuell auffälligen Eigenschaften erkannt (vgl. Gebhardt et al. 2016a, S. 7).

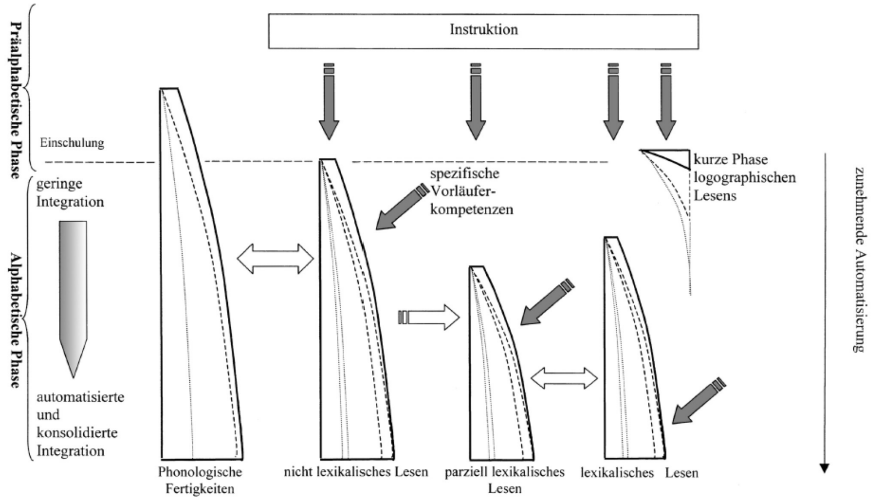


Abbildung 3: *Das Kompetenzentwicklungsmodell des Lesens (entnommen aus Klicpera et al. 2013, S. 32).*

Die nächste Phase, die alphabetische Phase mit geringer Integration, beinhaltet den allmählichen Aufbau für das Lesen notwendiger Kompetenzen. Somit beginnt hier der tatsächliche Leselernprozess. Zu den genannten für den Leseprozess notwendigen Kompetenzen gehören das alphabetische Prinzip sowie die phonologische Dekodierung. Allerdings sind diese Teilfähigkeiten noch nicht vollständig und fehlerfrei ausgebildet, sondern befinden sich zu diesem Zeitpunkt noch in der Ausbildung. Aufgrund der recht regelmäßigen Laut-Buchstaben-Zuordnungen des Deutschen können SuS diese Strategie bereits sehr schnell anwenden, sodass sie bereits nach einigen Wochen in der Schule fähig sind, Pseudowörter aus Graphemen, die die SuS bereits kennen, zu erlesen (vgl. Klicpera et al. 2013, S. 33). Durch die zunehmende Automatisierung der „Fähigkeit zum (schnellen) lexikalischen Abruf von Wörtern“ gelangen die SuS in die alphabetische Phase mit voller Integration. Dadurch lässt sich eine Abnahme an Fehlern und eine Steigerung der Lesegeschwindigkeit verzeichnen. Wie in der orthographischen Phase nach Frith und Günther lernen die SuS in dieser Stufe nicht mehr buchstabenweise zu dekodieren, sondern häufig vorkommende Buchstabenkombinationen wie Silben direkt zu erlesen. Diese Phase dient auch dazu zu automatisieren und zu verinnerlichen, welcher Lesezugang gewählt wird. Außerdem arbeiten beide Zugänge stetig mehr zusammen (vgl. ebd., S. 34).

3 Lernverlaufsdiagnostik

Im vorangegangen Kapitel wurde bereits die Bedeutung der Lesefähigkeit für den Bildungserfolg sowie für die gesamte gesellschaftliche Teilhabe angesprochen. Im Anschluss daran wird nun in Kapitel 3 die Relevanz der Lernverlaufsdiagnostik, die die Basis für eine optimale Förderung bietet, dargelegt. Weiterhin wird auf die Entstehung und die Konstruktion derartiger Verfahren eingegangen. Zur Vorarbeit für den folgenden empirischen Teil werden typische Lernverläufe dargestellt, ein theoretischer Input zu Boden- und Deckeneffekten gegeben sowie der aktuelle Forschungsstand zur Lernverlaufsdiagnostik im Lesen präsentiert.

3.1 Begründung der Lernverlaufsdiagnostik im (inklusiven) Unterricht

Der Einsatz eines Instrumentes, welches den Lernverlauf eines Schülers oder einer Schülerin misst, resultiert aus einer Vielzahl von Gründen. Dies bestätigt auch der von Blumenthal, Kuhlmann und Hartke (2014) aktuell wahrgenommene Wandel hin zu einer individuelleren und prozessorientierten Leistungsbewertung weg von der klassischen Statusdiagnostik (vgl. S. 42).

Zum einen hat die Lernverlaufsdiagnostik im Gegensatz zu üblicherweise angewandten Diagnoseinstrumenten den Vorteil, dass sie nicht nur einen punktuellen Leistungsstand feststellt, sondern über mehrere Testzeitpunkte hinweg, Lernfort- oder Rückschritte und auch mögliche Stagnationen erfasst (vgl. Klauer 2011, S. 207). Entgegen der früheren Annahme, dass „mit kleinen, klausurartigen Überprüfungen der Schülerleistungen […] die Produkte von Lehr-Lernprozessen objektiv gemessen, fair beurteilt und als Grundlage für gesellschaftlich notwendige Selektionsentscheidungen herangezogen werden“ können (Maier 2010, S. 294), wird heute die Auffassung vertreten, dass es notwendig ist, den Lernstand der SuS in regelmäßigen Abständen neu zu erfassen, da lediglich auf diese Weise eine adäquate Anpassung zwischen Lernstand der SuS und Förderung erreicht werden kann (vgl. Diehl, Hartke & Knopp 2009, S. 122). Denn Unterricht kann sich nur dann als erfolgreich erweisen, wenn er optimal an das Vorwissen der SuS anknüpft (vgl. Maier 2010, S. 299) und damit dem häufig genannten Leitsatz „Die Schüler da abholen, wo sie stehen“ gefolgt wird (vgl. Diehl et al. 2009, S. 122).

Zum anderen ist die Lernverlaufsdiagnostik besonders im Bereich des inklusiven Bildungssystems von hoher Relevanz. Aufgrund des Inkrafttretens der UN-Behindertenrechtskonvention im Jahr 2009 und insbesondere aufgrund des darin enthaltenden Artikels 24 sind alle Vertragsstaaten und somit auch Deutschland dazu verpflichtet, ein inklusives Bildungssystem zu etablieren (vgl. Deutsches Institut für Menschenrechte 2006, S. 15). Das deutsche Schulsystem steht seitdem vor der großen Herausforderung, die damit implizierten umfassenden

Veränderungen regelkonform umzusetzen. Eine immense Herausforderung - ungeachtet der finanziellen und bautechnischen Herausforderungen - und gleichzeitig eine große Chance, ist die durch die Inklusion stark ansteigende Heterogenität. Die ohnehin schon unterschiedlichen Lernausgangslagen werden durch den gemeinsamen Unterricht von SuS mit und ohne Förderbedarf noch vielfältiger. Derartige Lerngruppen benötigen daher ein hohes Maß an Individualisierung, da man den Bedürfnissen aller SuS gerecht werden muss. Um dies zu ermöglichen, müssen die Lernstände und Förderbedarfe der SuS regelmäßig ermittelt und auf Grundlage dessen entsprechende Fördermaßnahmen entwickelt werden (vgl. Mühling, Gebhardt & Diehl 2017, S. 556). Formative Leistungsmessungen, wie die Lernverlaufsdiagnostik, können diesen Anforderungen, aufgrund ihrer durch die wiederholten Testungen entstehenden Nähe zum Entwicklungsprozess, eher gerecht werden als die traditionelle Statusdiagnostik (vgl. Wilbert & Linnemann 2011, S. 225).

Trotz der Kritik, resultierend aus negativen Effekten auf die Schülermotivation und das Fähigkeitsselbstkonzept, werden in deutschen Schulen dennoch hauptsächlich die klassischen auf Notenvergabe und Selektion ausgerichteten Verfahren der Statusdiagnostik angewendet (vgl. Maier 2010, S. 293). Dabei können Verfahren zur Lernverlaufsdiagnostik für alle SuS und insbesondere für SuS mit Lernschwierigkeiten einen großen Vorteil bedeuten. Denn bereits bestehende oder sich andeutende Lernschwierigkeiten, wie beispielsweise Lese-Rechtschreib-Störungen, können mithilfe der Lernverlaufsdiagnostik frühzeitig erkannt werden, sodass eine intervenierende oder präventive Förderung einsetzen kann (vgl. Diehl 2011, S. 166). Allein auf diese Weise kann erreicht werden, dass sich Leseschwierigkeiten nicht dauerhaft einstellen (vgl. Jungjohann et al. 2018a, S. 101). Die besondere Unterstützung dieser Schülergruppe ist von sehr hoher Relevanz, da sich aufgrund häufiger Misserfolgserlebnisse ein negatives Fähigkeitsselbstkonzept und auch Versagensängste ausbilden können. In der Folge treten soziale Phobien, depressive Störungen sowie Störungen des Sozialverhaltens bei SuS mit Lesestörung deutlich häufiger auf als bei SuS, die keine Lesestörung zeigen, wodurch eine optimierte Leseförderung einen noch höheren Stellenwert erlangt (vgl. Galuschka & Schulte-Körne 2015, S. 474). Im Hinblick auf die Misserfolgserlebnisse und das negative Fähigkeitsselbstkonzept ist ein weiterer entscheidender Vorteil der Lernverlaufsdiagnostik, dass den SuS ihr individueller Lernstand und ihr individueller Lernverlauf zurückgespiegelt werden kann ohne sie an der sozialen Bezugsnorm zu messen (vgl. Gebhardt et al. 2016b, S. 444). Lernrückstände frühzeitig aufzudecken und die weiteren Lernprozesse daran anzupassen, ist ganz im Sinne des „Response-to-Intervention“ (RTI) -Ansatzes (vgl. Jungjohann et al. 2018a, S. 101). RTI-Strukturen wurden entwickelt, um gegen die sogenannte

„wait-to-fail“-Problematik zu arbeiten. Das wait-to-fail-Prinzip meint, dass „das deutsche Bildungssystem […] darauf ausgerichtet [ist], dass die Probleme eines Schulkinds umfassend und massiv werden müssen, bis sie mit den zur Verfügung stehenden Diagnoseinstrumenten zweifelsfrei erfasst und klassifiziert werden können“ (Huber & Grosche 2012, S. 313). Der RTI-Ansatz ist hingegen darauf ausgelegt, dass er gezielt überprüft, inwiefern SuS auf die verwendeten Fördermaßnahmen ansprechen, sodass diese bei zu geringem oder ausbleibendem Leistungszuwachs individuell auf die SuS zugeschnitten werden (vgl. Klauer 2014, S. 6).

Abgesehen von der Schülerperspektive kann die Lernverlaufsdiagnostik außerdem auch einen positiven Beitrag für die Lehrkräfte leisten. Denn durch die Begutachtung der Lernverläufe können sie vereinfacht Rückschlüsse auf den Erfolg ihres Unterrichts ziehen und in Abhängigkeit der Ergebnisse ihre angewandten Fördermaßnahmen stetig verbessern und an die Lernstände der SuS anpassen (vgl. Gebhardt et al. 2016b, S. 445). Inventare zur Messung des Lernverlaufs benötigen kein tiefergehendes Wissen und auch keine aufwändige Einarbeitung (vgl. Voß & Hartke 2014, S. 87). Dies ist von besonderer Relevanz, da die PISA-Studien neben den gravierenden Defiziten in der Lesekompetenz der SuS auch Defizite in der Diagnosekompetenz von den Lehrpersonen aufdecken konnte (vgl. Diehl 2011, S. 164). Selbsterklärend können einige Lehrkräfte die Leistungsstände allein aufgrund ihrer langjährigen Berufserfahrungen und ohne Nutzen spezieller Verfahren der Lernverlaufsdiagnostik oder ähnlichem korrekt diagnostizieren. Da jedoch psychologische Effekte (z.B. Beurteilungsfehler wie der Halo-Effekt oder der Strenge-/Mildeeffekt), die den Lehrkräften nicht immer bewusst sind, einen großen Einfluss auf die Beurteilung der Schülerleistungen haben, stellt dies ein weiteres Argument für die Etablierung der Lernverlaufsdiagnostik als objektives Diagnoseinstrumente für die Unterrichtspraxis dar (vgl. Diehl et al. 2009, S. 122).

3.2 Theoretischer Hintergrund der Lernverlaufsdiagnostik

Der Ursprung der Lernverlaufsdiagnostik geht zurück auf Stanley Deno, der seit 1972 mit einem Forscherteam an der Universität Minnesota in Minneapolis Interventionsprogramme zur Unterstützung der sonderpädagogischen Praxis erforschte (vgl. Klauer 2011, S. 207f.). „The primary goal of the research program was to develop measurement and evaluation procedures that teachers could use routinely to make decisions about whether and when to modify a student's instructional program“ (Deno 1985, S. 221). Mit der Entwicklung des Curriculum-Based Measurements (CBM) hat die Forschergruppe um Deno dieses Ziel 1983 erreicht. Das Verfahren des CBM zielte zu seinen Anfängen insbesondere auf die Erfassung von Lese-

und Schreibfähigkeiten von SuS mit Lernschwierigkeiten ab. Deno bezeichnete sein Diagnoseinstrument zur Lernverlaufsmessung als curriculum-basiert, da die Testungen genau die Kompetenzen messen sollten, die im Unterricht zu dem Zeitpunkt der Testungen auch wirklich vermittelt werden. So steht im Kern des CBM die Messung der Veränderungen von aktuell vermittelten Fähigkeiten über einen bestimmten Zeitraum hinweg (vgl. Deno 2003, S. 184).

Lange Zeit wurde dem CBM in Amerika nur wenig Beachtung geschenkt. Erst durch das Inkrafttreten des damaligen „No Child Left Behind"-Bildungsgesetzes von 2002, durch welches Schulen für Misserfolge ihrer SuS selbst verantwortlich gemacht wurden und bei häufigen Misserfolgen sogar mit Sanktionen zu rechnen hatten, gewann das CBM an Bedeutung. Denn CBM erwies sich als eine geeignete Methode, um Leistungsschwächen frühzeitig zu erfassen sowie um Lehrziele und den Unterricht an die individuellen Messergebnisse der SuS anzupassen (vgl. Klauer 2014, S. 3). „It is not surprising, but important to note that those teachers using CBM in formative evaluation were more accurate in identifying their students' goals" (Deno 2003, S. 187).

CBM zeichnet sich also dadurch aus, dass nicht lediglich eine einmalige Fähigkeitsüberprüfung stattfindet, sondern dass kontinuierlich erhobene Leistungstests durchgeführt werden, sodass der Lernverlauf über die Zeit hinweg grafisch dargestellt und betrachtet werden kann und die Lehrkraft die Unterrichtsprozesse optimal an diesen anpassen kann (vgl. Voß & Hartke 2014, S. 84).

Die Curriculum-Based Measurements umfassen, wie bereits beschrieben, Diagnoseinventare, die aktuell im Unterricht behandelte Kompetenzen, abfragen. Der Terminus soll in Abgrenzung zu anderen Maßnahmen des Curriculum-Based Assessments (CBA) außerdem betonen, dass CBM auf wissenschaftlich begründete Messungen abzielt, weswegen Deno auch die Begrifflichkeit „measurement" wählte (vgl. Klauer 2006, S. 17). Das CBA ist ein Oberbegriff für jegliche Maßnahmen, die zur Gewinnung von Informationen für die pädagogische Entscheidungsfindung dienen. CBA zeichnet sich durch die Curriculum-Nähe sowie durch die direkte Beobachtung von Schülerleistungen aus und gilt somit als fairer als standardisierte überregionale Schulleistungstests, die den im Unterricht tatsächlich vermittelten Inhalten nur wenig Beachtung schenken (vgl. Diehl et al. 2009, S. 124). CBM ist demnach als ein Teilbereich der CBA zu betrachten, der sich aufgrund des Bezugs zu wissenschaftlichen Standards, beispielsweise durch die Berücksichtigung der traditionellen Gütekriterien und aufgrund der Messung des Lernverlaufes, aber von anderen Teilbereichen der CBA abhebt (vgl. Voß und Hartke 2014, S. 85).

In Deutschland fand ein derartiges Diagnoseinventar, ähnlich wie in den Vereinigten Staaten, jahrzehntelang keine Beachtung (vgl. Klauer 2011, S. 208). Erst im Jahr 2006, mit der Veröffentlichung eines Zeitschriftenartikels von Klauer, begann die Aufmerksamkeit der deutschsprachigen sonderpädagogischen Forschung an den CBM-Verfahren stetig zu wachsen (vgl. Klauer 2006, S. 16, Klauer 2011, S. 208). Die weitere Entwicklung des CBM in Deutschland wurde maßgeblich von Jürgen Walter geprägt, da er bereits im Jahr 2010 ein erstes Testverfahren zur Leseflüssigkeit und im Jahr 2013 ein weiteres zum sinnerfassenden Lesen entwickelte (vgl. Klauer 2014, S. 7). Der Terminus des Curriculum-Based Measurements ist im deutschsprachigen Raum, auch im sonderpädagogischen Kontext, jedoch immer noch relativ unbekannt. Anstelle dessen wird hier häufig die Bezeichnung der Lernverlaufsdiagnostik verwendet.

Der Terminus der Lernverlaufsdiagnostik stellt sich bei genauerer Betrachtung als noch ein relativ junger Begriff heraus. Denn Klauer, der den Grundstein für die Lernverlaufsdiagnostik in Deutschland legte, prägte zunächst den Begriff der Lernfortschrittsmessung, da er diesen für besser geeignet hielt als eine wörtliche Übersetzung der CBM-Bezeichnung (vgl. Klauer 2006). In Anlehnung daran benannte Walter seine Testverfahren zur Messung der Leseflüssigkeit auch als „Lernfortschrittsdiagnostik Lesen“ (vgl. Walter 2010). Infolge von ersten Erprobungen der neuen Diagnoseinventare und durch die Erkenntnis, dass die Lernverläufe der SuS nicht nur durch Lernfortschritte, sondern auch durch Rückschritte und Stillstände gekennzeichnet sind, wurde der Begriff der Lernfortschrittsdiagnostik schnell als unpassend abgelegt und der Terminus der Lernverlaufsdiagnostik letztendlich etabliert (vgl. Klauer 2011, S. 208). Ähnlich wie die in Deutschland bekannte Förderdiagnostik soll die Lernverlaufsdiagnostik nicht im Sinne der Statusdiagnostik zur Selektion oder Klassifikation aufgrund der erbrachten Leistungen führen. Denn Lernverlaufsdiagnostik meint lediglich die „Dokumentation des Lernfortschritts im Verlauf der Zeit“ (Klauer 2006, S. 17).

Eine weitere begriffliche Unterscheidung, die im Hinblick auf die Lernverlaufsdiagnostik erwähnt werden sollte, ist die Abgrenzung der Begriffe des summativen und des formativen Assessments. Die Unterscheidung dieser Begrifflichkeiten geht zurück auf Scriven, der diese Bezeichnungen bereits im Jahr 1967 prägte (vgl. Scriven 1967, S. 43). Während die summative Leistungsmessung oder auch summative Evaluation keinen neuen Ansatz bereithält, da sie die Leistungsmessung am Ende einer Lernperiode oder am Ende eines Schuljahres bezeichnet und demnach z. B. die Notenvergabe oder Selektionsentscheidungen durch klassische schriftliche Klausuren umfasst, stellt das formative Assessment einen neuen Ansatz dar (vgl. Klauer 2014, S. 1f.). Der Terminus der formativen Leistungsmessung ist in Deutschland

eher unbekannt. Konzepte, die unter das formative Assessment fallen, werden in der deutschen Schulpraxis häufiger als alternative Leistungsmessungen bezeichnet und meinen Messformen wie handlungsorientierte Aufgaben, Portfolios o.ä. (vgl. Maier 2010, S. 294f.). Der Begriff der formativen Evaluation ist ihrem Ursprung nach jedoch umfassender. Er bezeichnet in regelmäßigen Abständen wiederholt stattfindende Leistungsmessungen und Bewertungen der Schülerleistungen sowie die fortlaufende Anpassung der Lehr-Lernprozesse an die erfassten Schülerleistungen (vgl. Maier 2017, S. 194). Die fortlaufenden Testungen sowie die ständige Optimierung der Lehr-Lernprozesse auf Basis der ermittelten Schülerleistungen stellen somit einen gemeinsamen Kern des curriculumbasierten Messens und der formativen Leistungsmessung dar, sodass CBM auch als eine spezielle Art der formativen Leistungsmessung bezeichnet wird (vgl. Klauer 2014, S. 4).

Hinsichtlich einer genauen Abgrenzung dieser Begrifflichkeiten kommt die Forschungsliteratur jedoch zu unterschiedlichen Resultaten. Denn während Balt, Ehlert und Fritz (2017) zu dem Entschluss kommen, dass sie synonyme Verwendung finden (vgl. S. 166), erklären Wilbert und Linnemann (2011), dass „unter den Begriffen Lernfortschrittsmessung, lernprozessbegleitende Diagnostik, Dynamic Testing, Response-to-Intervention oder Curriculum-basierte Messung im Detail zwar unterschiedliche, aber eng verwandte Konzepte diskutiert“ werden (S. 226). Dennoch erkennen sie an, dass der gemeinsame Fokus der Konzepte in der Betrachtung unterschiedlicher Lernverläufe liegt und verwenden daher weiterhin die Bezeichnung der Lernverlaufsdiagnostik (vgl. ebd.). Diesem Vorbild wird auch im Rahmen dieser Ausführungen gefolgt.

Diagnoseverfahren dienen im Allgemeinen der Messung eines bestimmten Merkmals oder einer bestimmten Fähigkeit (vgl. Wilbert & Linnemann 2011, S. 226). Lernverlaufsdiagnostik bedeutet demgemäß, dass Lernzuwächse einer bestimmten Kompetenz über eine gewisse Zeitspanne mehrmals gemessen werden. Durch die wiederholten Messungen der Lernstände kann dokumentiert werden, wie sich die spezifische Kompetenz der SuS in dieser Zeitspanne entwickelt (vgl. Klauer 2006, S. 16f.). Damit die Erhebungen auch tatsächlich wiederholt stattfinden, sollen die Testaufgaben effizient und möglichst kurz gestaltet sein. Um die Praxistauglichkeit weiter zu erhöhen, sind nach Deno Effizienz, ein geringer Kostenaufwand und leichte Verständlichkeit auch in Bezug auf die Testergebnisse, sodass Ergebnisse leicht an Eltern, Lehrer und an die SuS kommuniziert werden können, weitere unbedingte Eigenschaften der Lernverlaufsdiagnostik (vgl. Deno 1985. S. 221). Bezüglich der Rückmeldung an die SuS wird betont, dass nicht nur die soziale Bezugsnorm, sondern auch die curriculare und vor allem die individuelle Bezugsnorm als Vergleichsmaßstab hinzugezogen werden kann (vgl. ebd.,

S. 230). Denn da für alle SuS ein individueller Lernverlauf ermittelt wird, können Schülerleistungen mittels der Lernverlaufsdiagnostik auch anhand der individuellen Bezugsnormen begutachtet werden und dementsprechende Rückmeldungen an die SuS zurückgegeben werden (vgl. Förster, Kuhn & Souvignier 2017, S. 117). Fällt auf, dass Lernverläufe stagnieren oder rückläufig werden, kann aufgrund der Lernverlaufsmessung frühzeitig interveniert und auf andere oder weitere Fördermaßnahmen zurückgegriffen werden (vgl. Klauer 2006, S. 17). Durch den Vergleich mit der sozialen Norm kann der individuelle Lernverlauf auch in Bezug zur Klassen- oder Jahrgangsstufenleistung gesetzt werden (vgl. Förster, Kuhn & Souvignier 2017, S. 117).

Instrumente der Lernverlaufsdiagnostik müssen bestimmten Anforderungen gerecht werden. Neben einer zeitökonomischen Gestaltung, der Berücksichtigung der drei klassischen sowie weiteren Gütekriterien, zeichnen sich derartige Verfahren dadurch aus, dass sie den Teilnehmern eine unmittelbare Ergebnisrückmeldung liefern sowie, dass sie sich als anschlussfähig für folgende individuelle Fördermaßnahmen erweisen sollen. Außerdem sollten die Lernverlaufsinstrumente eine breite Schwierigkeitsstreuung aufweisen, um die Leistungen aller SuS abzubilden und Boden- und Deckeneffekte möglichst zu vermeiden (vgl. Souvignier, Förster und Salaschek 2014b, S. 241).

Im deutschsprachigen Raum liegen, trotz der Zunahme in den letzten Jahren, immer noch erst wenige Forschungsergebnisse vor. Die bereits vorhandenen zeigen jedoch, dass die Konzepte auch in Deutschland erfolgreich einsetzbar sind (vgl. Voß & Hartke 2014, S. 95). In den letzten Jahren wurden in Deutschland insbesondere Lernverlaufsmessungen für den Bereich der Lesefähigkeit, des Schreibens und des Rechnens entwickelt. Dies meint jedoch nicht, dass Instrumente zur Lernverlaufsdiagnostik nicht auch für weitere Unterrichtsfächer entwickelt werden können (vgl. Walter 2009, S. 164). Bei den in Deutschland entwickelten Tests kann jedoch eigentlich nicht mehr von CBM im Sinne einer am Curriculum orientierten Messung gesprochen werden, da beispielsweise in Lernverlaufsmessungen zur Lesefähigkeit grundlegende Teilkompetenzen übergeprüft werden, die nicht dem entsprechen, was aktuell im Unterricht vermittelt wird (vgl. Klauer 2014, S. 5). Auch da die Lernverlaufsdiagnostik bestenfalls über einen längeren Zeitraum stattfindet, kann sie dem Anspruch, das zu messen, was aktuell im Unterricht gelernt wird, nicht gerecht werden, sodass häufiger das gemessen wird, was bis zum Ende eines Schuljahres erlernt werden soll (vgl. Klauer 2011, S. 211). Zu diesem Resultat kamen auch Jungjohann et al. (2018a) nach einem systematischen Review von acht formativen Lernverlaufsinstrumenten, von denen keins die Möglichkeit enthielte „die vorgegebenen Inhalte an das Curriculum“ oder sogar! „an den individuellen Lernstand der Kinder anzupassen“ (S. 110).

Klauer (2011) unterscheidet zwei verschiedene Formen der Lernverlaufsdiagnostik. In der ersten Variante wird eine Kompetenz erfasst, die im Grunde genommen bereits von den SuS beherrscht wird, aber weiter geübt wird, um beispielsweise eine schnellere Aufgabenbewältigung oder eine geringere Fehleranzahl zu erreichen. In diesem Fall werden also Veränderungen der Kompetenzen hinsichtlich der Geschwindigkeit bzw. der Genauigkeit gemessen. Ein Beispiel dafür ist die Messung der Lesefertigkeit. Die zweite Form misst die Kompetenzen, die am Ende des Schuljahres erreicht werden sollen. Die Kompetenzen sollen an dieser Stelle also um neue Komponenten erweitert werden. Für die Lernverlaufsdiagnostik ergibt sich daraus aber ein entscheidender Nachteil. Denn, da alle Tests gleich schwer sein müssen und immer die gleiche Fähigkeit gemessen werden muss, müssen bei dieser zweiten Variante bereits Aufgaben enthalten sein, die die SuS mit ihrem Kenntnisstand zu Beginn der Lernverlaufsmessung noch nicht lösen können (vgl. S. 219f.).

Auch Walter (2008) konnte mit seiner Studie zur Lernverlaufsdiagnostik im Lesen beweisen, dass CBM nicht nur in Amerika positive Effekte erzielt, sondern auch im deutschsprachigen Raum wissenschaftlichen Überprüfungen zur Validität, Reliabilität und Änderungssensibilität standhalten kann (vgl. S. 77). Klauer (2006) weist jedoch auch daraufhin, dass trotz der erfreulichen Ergebnisse bezüglich der Reliabilität und der Validität immer mit Messfehlern zu rechnen ist und auch Tagesschwankungen ein verfälschtes Bild ergeben können, sodass diesen am besten durch ein häufiges Testen vorgebeugt wird (vgl. S. 25). CBM stellt jedoch nur eine Messtechnik dar und zum Leistungszuwachs der SuS gehört mehr als lediglich die Diagnose über den Lernverlauf (vgl. Walter 2008, S. 77; Gebhardt, Heine, Zeuch, & Förster, 2015). So erklärt Walter (2010) weiterhin, dass „Lehrer, die ihren Unterricht mithilfe von CBM einer systematischen formativen Evaluation unterziehen, […] bei ihren Schülern einen Lernzuwachs von etwa einer Dreiviertel Standardabweichung (ES = .70) oder mehr erzeugen“ können, aber nur insofern sie zusätzlich Techniken der Verhaltensmodifikation gebrauchen, sich an die Regeln zur Dateninterpretation halten und auch die visualisierten Lernverläufe nutzen (S. 163). Auch Souvignier, Förster & Schulte (2014a) kommen zu dem Resultat, dass die Effektivität von formativen Leistungsmessungen nicht grundsätzlich vorhanden ist, sondern von bestimmten Bedingungen abhängt (vgl. S. 221). Folglich entsteht durch die alleinige Lernverlaufsmessung kein Lernzuwachs, was bedeutet, das CBM lediglich als ein „Impulsgeber in einem mehr oder minder qualifiziert stattfindenden Vermittlungsprozess“ fungiert (Diehl et al. 2009, S. 125).

Formative Assessments können, trotz der Vorteile gegenüber summativen und selektierenden Ansätzen, diese nicht ersetzen. Denn allein aufgrund der formativen Leistungsbeurteilung werden nicht alle Funktionen der klassischen Leistungsbeurteilung, wie beispielsweise

die Selektionsentscheidung, erfüllt (vgl. Bürgermeister et al. 2014, S. 46). Außerdem ermöglicht die Lernverlaufsdiagnostik lediglich die Darstellung der Lernverläufe, gibt aber keine Hinweise auf die Ursachen möglicher Lernschwierigkeiten (vgl. Klauer 2006, S. 25).

Trotz der genannten Kritikpunkte liegt der entscheidende Vorteil der Lernverlaufsdiagnostik darin, dass eine frühzeitige Erkennung von Lernschwierigkeiten das Fundament für eine optimale Anpassung des Unterrichts an die individuellen Lernbedingungen der SuS darstellt (vgl. Diehl 2011, S. 171). Durch die wiederholten Messungen kann es ermöglicht werden, dass SuS eine sonderpädagogische Unterstützung bekommen ohne bereits schwerwiegende Lernprobleme und Rückstände aufzuweisen (vgl. Jungjohann et al. 2018a, S. 101) und somit dem Vorwurf, dass das „das deutsche Bildungssystem […] darauf ausgerichtet [ist], dass die Probleme eines Schulkinds umfassend und massiv werden müssen, bis sie mit den zur Verfügung stehenden Diagnoseinstrumenten zweifelsfrei erfasst und klassifiziert werden können“ entgegen gearbeitet wird (Huber & Grosche 2012, S. 313). Daher profitieren zwar insbesondere SuS mit Lernschwierigkeiten von der Lernverlaufsdiagnostik, insgesamt betrachtet können bei optimierten Lernprozessen, infolge der Lernverlaufsdiagnostik, aber alle SuS ihren Vorteil ziehen (vgl. Maier 2010, S. 300). Denn erst durch die Lernverlaufsdiagnostik wird die Diskrepanz zwischen dem aktuellen Leistungsstand und dem erwarteten Lernziel offenbart und kann zur Überbrückung dieser beitragen (vgl. Blumenthal et al. 2014, S. 42). So kommt auch Wayman et al. (2007) zu dem Entschluss, „[that] it was necessary for teachers to have a tool that could be used to evaluate growth in response to instruction. CBM was developed to serve that purpose“ (S. 85).

Die Ursache, warum formative Assessments trotz jahrelanger Forschungen jedoch immer noch am Anfang stehen (vgl. Jungjohann et al. 2018a, S. 101), liegt wahrscheinlich darin, dass die praktische Umsetzung sowie die Konstruktion der Tests noch vor einigen Hindernissen stehen. Ein Hindernis der praktischen Durchführung begründet sich beispielsweise in dem erforderlichen Engagement der Lehrkräfte. Regelmäßig durchgeführte Testungen benötigen einerseits viel Unterrichtszeit, andererseits nehmen sie aber auch viel Zeit und Aufwand für die Korrektur und Auswertung der Tests in Anspruch. Forscher versuchen daher Lernverlaufsmessungen noch praktikabler zu gestalten. Eine Möglichkeit stellt beispielsweise die computerbasierte Diagnostik dar. Ein Exempel dafür ist das Lernverlaufsdiagnoseinventar quop (vgl. Souviginier & Förster 2011, S. 245) oder das im folgenden empirischen Teil dargestellte Diagnoseinstrument Levumi.

Weitere Schwierigkeiten bezüglich der Konstruktion von Instrumenten zur Lernverlaufsdiagnostik werden im anschließenden Kapitel ausführlich dargestellt.

3.3 Konstruktion der Tests

Die bereits angesprochenen Schwierigkeiten in der Konstruktion der Tests begründen sich unter anderem darin, dass es sich bei der Lernverlaufsdiagnostik um ein relativ neues und daher noch nicht vollständig erforschtes Verfahren handelt. Folglich konnten einige Schwierigkeiten bereits beseitigt werden, andere werden allerdings erst zukünftig noch gelöst werden müssen.

Ein erstes Hindernis bei der Konstruktion von Tests zur Lernverlaufsdiagnostik ist die zugrundeliegende Testtheorie, da sich nicht auf die klassische Testtheorie (KTT) berufen werden kann. Der Hauptgrund dafür liegt nach Balt et al. (2017) darin, dass Grundannahmen der KTT, wie beispielsweise das Intervallskalenniveau der gemessenen Skala, nicht ausreichend überprüft werden können. Da in der Lernverlaufsdiagnostik wiederholte Tests zur Erfassung von Veränderungen durchgeführt werden, stellt sich dies problematischer dar als bei einmalig durchgeführten Diagnoseverfahren (vgl. ebd., S. 167). Klauer (2014) führt darüber hinaus an, dass „mit klassisch konstruierten Tests das Reliabilitäts-Validitäts-Dilemma“ nicht überwunden werden kann (S. 11). Denn hohe Parallel- und Retestreliabilitäten lassen sich nicht mit den zu ermittelnden Veränderungen durch die Lernverlaufsdiagnostik vereinbaren. Da Tests einer Lernverlaufsdiagnostik änderungssensibel sein müssen, sollten sie zwar hohe Werte der Split-half Reliabilität aufweisen, dementgegen jedoch nur mäßige Werte der Parallel- und Retestreliabilitäten (vgl. ebd., S. 11). Messinstrumente zur Lernverlaufsdiagnostik sollten also das Kriterium der Änderungssensiblität erfüllen. „To be most useful, the procedures should be sensitive to growth in student performance over relatively short durations“ (Deno 1985, S. 225). Änderungssensibilität bedeutet folglich, dass der Test bereits geringe Veränderungen auch innerhalb eines kurzen Zeitraums zuverlässig erfassen kann. In der Lernverlaufsdiagnostik genügt es aber nicht änderungssensible Items zu entwickeln, denn da die Veränderungen über eine gewisse Zeitspanne erfasst werden sollen, müssen alle Tests änderungssensitiv sein (vgl. Klauer 2011, S. 212). Dies wird unter anderem, wie bereits beschrieben, durch eine mäßige Paralleltest- und Retestreliabilität erreicht (vgl. Klauer 2014, S. 171). Außerdem gilt auch die Eindimensionalität, d.h. der Fokus auf eine spezifische Kompetenz, als Voraussetzung dafür, dass ein Test änderungssensibel misst (vgl. Wilbert & Linnemann 2011, S. 227).

Darüber hinaus fokussiert die KTT bei der Konstruktion änderungssensitiver Tests die Parameter der Itemschwierigkeiten und der Itemtrennschärfe. Dabei sollen die Itemschwierigkeiten bestenfalls über einen möglichst großen Kompetenzbereich streuen und die Trennschärfeindizes ebenfalls größtmöglich sein (vgl. Klauer 2011, S. 211). Enthält der Test nun, wie es

die Lernverlaufsdiagnostik vorsieht und folgend noch ausführlich erklärt wird, immer neue Aufgaben, so ergibt es keinen Sinn diese Parameter noch zu ermitteln, wodurch jedoch auch keine Paralleltestreliabilität ermittelt werden kann. Da die Tests durch immer neue Aufgaben gebildet werden, kann außerdem ebenso wenig die Retestreliabilität bestimmt werden. Des Weiteren wird auch die Berechnung von Cronbachs α sinnlos, da dafür jeder Proband dieselben Tests erhalten müsste. Dies ist in der Lernverlaufsdiagnostik jedoch ebenfalls nicht der Fall (vgl. Klauer 2014, S. 11). „Kurz und gut: Die Verfahren der Lernverlaufsdiagnostik gemäß der klassischen Testtheorie zu konstruieren ist praktisch ausgeschlossen“ (Klauer 2011, S. 211).

Anders sieht es in der probabilistischen Testtheorie (PTT) aus. Besonders Prozessmodelle der Item-Response-Theorie bieten sich für die Lernverlaufsdiagnostik an. In der PTT ist die Wahrscheinlichkeit für die Lösung eines Items eine Funktion aus dem Itemparameter und dem Fähigkeitsparameter des Probanden (vgl. Klauer 2011, S. 211f.). Demnach werden die gemessenen Werte nach dieser Theorie als manifeste Indikatoren einer latenten Personeneigenschaft (Fähigkeit), welche eigentlich erfasst werden soll, angesehen. Das Antwortverhalten der Probanden, d.h. die richtige oder falsche Lösung einer Aufgabe, ist von Personenmerkmalen (z. B. von den Fähigkeiten) und von Aufgabenmerkmalen (z. B. von der Schwierigkeit) abgängig. Durch die Annahme, dass die gemessenen Werte in der probabilistischen Theorie lediglich als manifester Indikator einer latenten Personeneigenschaft angesehen und durch das Rasch-Modell spezifiziert werden können und nicht wie in der KTT direkt als Fähigkeit der Person gilt, bietet sich die PTT für die Lernverlaufsdiagnostik an (vgl. Balt et al. 2017, S. 167f.).

Wilbert und Linnemann (2011) führen an, dass sich die Kriterien, denen die Lernverlaufsdiagnoseverfahren gerecht werden muss, prinzipiell nicht von denen der klassischen Statusdiagnostik unterscheiden. Die Kriterien müssen lediglich anders akzentuiert werden, wodurch aber Probleme in der praktischen Umsetzung resultieren können. Als zentrale testtheoretische Anforderungen führt er die Itemschwierigkeit, die Eindimensionalität, die Testfairness und die Reliabilität an (vgl. S. 226).

Das Kriterium der *Itemschwierigkeit* zu erfüllen, stellt bereits eine erste Hürde bei der Konstruktion eines Instruments zur Lernverlaufsdiagnostik dar. In der Lernverlaufsdiagnostik werden mehrmalig kurze Tests durchgeführt. Um den Lernverlauf zuverlässig abzubilden, müssen alle Tests immer auf dem gleichen Schwierigkeitsniveau angelegt sein. Diese sogenannte Homogenität der Testschwierigkeit ist für die Lernverlaufsdiagnostik von hoher Relevanz. Um dieses Kriterium zu erfüllen, dürfen jedoch nicht exakt dieselben Tests wiederholt

werden, denn aufgrund der Testwiederholung würde die Bearbeitung dessen für die SuS leichter und besser ausfallen und die Ergebnisse folglich verfälscht werden. Die erste Schwierigkeit der Testkonstruktion liegt daher darin, Tests zu entwickeln, die immer gleich schwer sind (vgl. Klauer 2011, S. 209f.). Denn nur wenn die Tests exakt denselben Schwierigkeitsgrad haben, können auftretende Veränderungen der Testergebnisse zwischen den Messungen als Veränderung der Fähigkeit des Probanden, d.h. als möglichen Lernzuwachs - denkbar sind natürlich auch Rückschritte und Stagnationen - gewertet werden (vgl. Wilbert & Linnemann 2011, S. 228). Da die Lernverlaufsmessung im besten Fall über einen längeren Zeitraum stattfindet und man die Annahme verfolgt, dass sich Leistungszuwächse ergeben, werden die Tests aufgrund des gleichbleibenden Schwierigkeitsniveaus und des zunehmenden Lernzuwachses mit der Zeit wahrscheinlich ohnehin leichter für die SuS (vgl. ebd., S. 210).

Das zweite Kriterium, welches bei der Testkonstruktion für Schwierigkeiten sorgen kann, ist die *Eindimensionalität*. Die Eindimensionalität umfasst die Testeigenschaft, dass „die gemessenen Leistungen in den Items eines Tests auf genau einen gemeinsamen latenten Faktor zurückzuführen sind“ (Wilbert & Linnemann 2011, S. 227). Denn um Veränderungen bezüglich einer bestimmten Fähigkeit durch die Lernverlaufsdiagnostik zu erfahren, müssen die Ergebnisse auch auf eben diese eine bestimmte Fähigkeit wieder zurückgeführt werden können. Das bedeutet, sollte beispielsweise ein Test die Lesekompetenz diagnostizieren, so sollte dieser Test auch allein durch die Lesekompetenz bestimmt sein und nicht von weiteren Kompetenzen abhängig sein (vgl. Wilbert 2014, S. 286). Folglich müssen die einzelnen Items der Lernverlaufstests nicht nur immer denselben Schwierigkeitsgrad aufweisen, sondern auch immer dasselbe Kompetenzspektrum abfragen (vgl. Klauer 2011, S. 209f.).

Ein weiteres Kriterium ist das der *Testfairness*. Ein Test gilt dann als fair gestaltet, wenn er „zu keiner systematischen Benachteiligung bestimmter Personen aufgrund ihrer Zugehörigkeit zu ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppen führt“ (Moosbrugger & Kelava 2008, S. 23). Eine mögliche Ursache für eine nicht vorhandene Testfairness liegt in dem Itembias. Unter der Bezeichnung des Itembias versteht man die Annahme, dass ein Item, welches aufgrund von nicht Test relevanten Merkmalen für unterschiedliche Personen unterschiedlich schwierig ausfallen kann. Ein Beispiel dafür: eine Schülerin mit einer anderen Erstsprache als Deutsch kann eine Mathematikaufgabe aufgrund unzureichender Deutschkenntnisse nicht lösen, obwohl ausreichend mathematische Kompetenzen vorhanden wären. Der Test dieses Beispiels erfüllt das Kriterium der Testfairness folglich nicht (vgl. Wilbert & Linnemann 2011, S. 229).

Die Einhaltung der drei klassischen Gütekriterien - *Reliabilität*, *Objektivität* und *Validität* - stellt ein weiteres Kriterium dar. Zu Beginn dieses Kapitels wurde bereits auf die Reliabilität in Lernverlaufsinstrumenten eingegangen, daher wird sie an dieser Stelle nicht weiter berücksichtigt und sich auf die Objektivität und Validität konzentriert. Ein Test kann als objektiv angesehen werden, wenn zwei unterschiedliche Rater denselben Test durchführen und übereinstimmende Testergebnisse erhalten. Testergebnisse sind also objektiv, wenn sie unabhängig vom Testdurchführer bestehen (vgl. Raithel 2008, S. 46). Um die Objektivität zu erhalten wird in Lernverlaufsmessungen häufig darauf hingewiesen, dass Anleitungen zur Durchführung und zur Auswertung wörtlich zu übernehmen sind, da so für die notwendige Durchführungs- und Auswertungsobjektivität gesorgt wird (vgl. Walter 2009, S. 166).

Die Validität, sprich die Gültigkeit von Lernverlaufsinstrumenten, gestaltet sich dementgegen etwas komplexer. Validität meint, ein Test misst genau das, was er auch beabsichtigt zu messen (vgl. Raithel 2008, S. 47). Da die Lernverlaufsdiagnostik ohnehin darauf abzielt, dass alle Tests den aktuellen Lernstand ein- und derselben Kompetenz ermitteln, so müssen sie folglich valide sein (vgl. Klauer 2011, S. 210). Messinstrumente zur Erfassung des Lernverlaufs müssen außerdem kontentvalide sein. Dies muss bereits in der Konstruktion der Tests erreicht werden. Denn erst wenn die Testaufgaben zu einer spezifischen Kompetenz eine repräsentative Stichprobe der Grundmenge aller Aufgaben dieser Kompetenz darstellen, gelten sie als kontentvalide. Um die Grundmenge der Aufgaben zu definieren, listet Klauer (2011) drei Möglichkeiten auf. Möglichkeit eins besteht darin eine Liste aller zu der Grundmenge zugehörigen Aufgaben anzufertigen. Anhand der zweiten Möglichkeit wird der bestimmte Sachverhalt durch Aussagesätze vollständig dargestellt, um diese Sätze anschließend in Testaufgaben umzuwandeln. Diese zweite Möglichkeit eignet sich im Besonderen für das Fach Sachunterricht. Die dritte Variante definiert die Grundmenge durch eine Aussageform. Das bedeutet, dass neben einer Konstanten die Aussage auch mindestens eine Variable enthält. Diese Möglichkeit bietet sich besonders für die Mathematik an. Da sich die Lernverlaufsdiagnostik dadurch auszeichnet, dass jeder Proband eine individuelle und bei jeder Testung eine neue Stichprobe von Aufgaben erhält, muss jeder dieser vielzähligen Tests die Grundgesamtheit valide abbilden. Sobald die Grundgesamtheit angemessen definiert wurde, kann eine repräsentative Itemstichprobe dieser Gesamtheit hergestellt werden. Zum einen kann, insofern sich die Menge der Aufgaben in sich als homogen erweisen, eine Zufallsstichprobe der Aufgaben gezogen werden. Zum anderen muss, wenn keine Homogenität vorliegt, ein proportional-zufälliges Verfahren angewendet werden, indem festgelegt wird, wie groß die Anteile der unterschiedlichen Teilmengen jeweils sein sollen, bevor ebenfalls durch einen Zufallsgenerator eine Zufallsstichprobe gezogen wird (vgl. Klauer 2011, S. 213f.).

Zur Auswahl der Items für einen Test formuliert Fuchs (2004) zwei unterschiedliche Herangehensweisen, einerseits den Ansatz des „*robust indicators*“ und andererseits den des „*curriculum samplings*“ (vgl. S. 189). Innerhalb des robust-indicator-Ansatzes „werden Kompetenzen bzw. Aufgaben ausgewählt, die sich empirisch als prädiktiv valide für die Gesamtleistung im interessierenden Bereich erwiesen haben“ (Schwenk, Kuhn, Doebler & Holling 2017, S. 125). Um Aufgaben zu entwickeln, die also spezifische Teilaspekte einer Kernkompetenz repräsentieren, muss diese Kompetenz zur Operationalisierung theoretisch fundiert werden. Dadurch können die Aufgaben beispielsweise in Zusammenhang mit einem Entwicklungsmodell gesetzt werden, um Auskunft über den fortschreitenden Entwicklungsprozess zu erhalten. Durch die Überprüfung der Zusammenhänge mittels der PTT werden auch qualitative Aussagen über den derzeitigen Entwicklungsstand der SuS möglich (vgl. Balt et al. 2017, S. 168). Der robust-indicator-Ansatz ermöglicht eine hohe Flexibilität, da er über mehrere Jahrgangsstufen hinweg eingesetzt werden kann. Im Gegensatz zum Ansatz des curriculum samplings erfolgt die Auswahl mittels des robust-indicator-Ansatzes eher induktiv (vgl. Schwenk et al. 2017, S. 125). Curriculum sampling meint, dass Stichproben aus einem Aufgabenpool gezogen werden. Dieser Pool besteht aus Aufgaben, die alle zu erlernenden Kompetenzen eines Schuljahres umfassen. Gegenüber dem robust indicator kann das curriculum sampling daher nicht über mehrere Jahrgangsstufen hinweg eingesetzt werden, bezieht sich dafür aber auf ein spezifisches Curriculum (vgl. Voß 2013, S. 199f.). Auch qualitative Aussagen über den Lernstand der SuS sind beim curriculum sampling nicht möglich (vgl. Balt et al. 2017, S. 168).

Lord und Novick haben bereits im Jahr 1968 festgestellt, dass es sich um ein *Binomialmodell* handelt, wenn jeder Teilnehmer seine eigene Zufallsstichprobe von Aufgaben erhält (S. S. 523f.). Das Binomialmodell geht davon aus, dass jeder Teilnehmer den Fähigkeitsparameter p besitzt, der darüber entscheidet, die Aufgaben zu lösen (vgl. Klauer 2011, S. 215). Die Aufgaben können gemäß einer dichotomen Datenstruktur entweder als richtig (= 1) oder als falsch (= 0) bewertet werden ($0<p<1$) (vgl. Balt et al. 2017, S. 174). Das Binomialmodell kann jedoch nur unter der Berücksichtigung zweier Voraussetzungen angewendet werden. Zum einen müssen alle Testaufgaben denselben Schwierigkeitsgrad aufweisen, zum anderen muss eben jeder Proband seine eigene Zufallsstichprobe erhalten. Im Sinne des Binomialmodells können nach Lord und Novick die Zufallsstichproben auch als Paralleltests aufgefasst werden, sodass auf diese Weise auch die Paralleltest-Reliabilität ermittelt werden kann (vgl. Klauer 2011, S. 215f.).

Auch die Normierung von Instrumenten zur Lernverlaufsdiagnostik stellt andere Bedingungen als die Statusdiagnostik. Die Lernverlaufsdiagnostik fordert somit Normen, die die Kompetenzveränderung innerhalb eines bestimmten Zeitraums umfassen. Förster, Kuhn und Souvignier (2017) weisen darauf hin, dass die Normierungen je nach Einsatzzweck individuell angepasst werden müssen, geben aber einen Überblick über unbedingt zu beachtende Aspekte (vgl. S. 116f.). So sollten zum einen bereits die Stichprobe, der Unterricht sowie die spezifische Testsituation repräsentativ sein und konkret beschrieben werden. Relevant ist an dieser Stelle zum Beispiel, dass die Lernausgangslagen der SuS dargelegt werden, denn der weitere Lernverlauf kann sich je nach Lernausgangslage unterscheiden. Auch die Anzahl der Messzeitpunkte sowie die Länge des Zeitraums müssen gut gewählt werden (vgl. ebd., S. 118). Von Bedeutung ist außerdem der Zeitpunkt der Messung; insbesondere im Hinblick auf mögliche Ferieneffekte. Denn beispielsweise kam eine Studie von Fink et al. (2015) zu dem Ergebnis, dass Lesefähigkeiten über die Ferien zwar gesteigert werden, mathematische und Rechtschreibkompetenzen sowie die Intelligenz sich jedoch verschlechtern. Insgesamt sollte allerdings auch beachtet werden, dass Forschungen zu den Ferieneffekten widersprüchlich sind und auch Ergebnisse existieren, in denen sich Ferieneffekte nachteilig auf die Lesekompetenz auswirkten (vgl. S. 311f.).

3.4 Stereotype Muster in den Lernverläufen

Obwohl die mittels formativen Assessments erhobenen Lernverläufe individuell für alle SuS oder für einzelne Klassen oder ganze Jahrgangsstufen betrachtet werden sollten, lassen sich einige Regelmäßigkeiten in den Verläufen erkennen.

Die ermittelten Lernverläufe der SuS werden als Liniendiagramme, d.h. in einem Koordinatensystem dargestellt. Dabei werden an der horizontalen x-Achse die verschiedenen Messzeitpunkte eingetragen, während die y-Achse beispielsweise die Anzahl der korrekt gelesenen Wörter enthält (vgl. Ardoin et al. 2013, S. 2). Enthält ein Diagramm mindestens vier Messzeitpunkte, wird aufgrund visueller Beurteilung entschieden, welche Verlaufsform die Daten optimal repräsentiert (vgl. Förster et al. 2017, S. 119). Meist ist eine lineare Regression zur Beschreibung des Lernverlaufs ausreichend. Lediglich wenn die Lernverlaufskurve eine negative oder positive Beschleunigung aufweist, muss auf Polynome zweiten oder höheren Grades zurückgegriffen werden. Dies tritt jedoch deutlich seltener auf (vgl. Strathmann & Klauer 2010, S. 119). Denn quadratische Verläufe sind ein Anzeichen dafür, dass die Lernzuwächse abnehmen (vgl. Förster et al. 2017, S. 119).

Lernverlaufskurven, die den Mittelwert des Zuwachses einer gesamten Klasse darstellen, unterscheiden sich meist sehr deutlich von den Verläufen einzelner SuS. Während der Lernverlauf der gesamten Klasse meist durch einen linearen Anstieg, d.h. durch einen stetig zunehmenden Leistungsanstieg gekennzeichnet ist, weisen die individuellen Verläufe der Probanden mehr Variabilität auf. Lernverlaufskurven können sich hinsichtlich des Ausgangsniveaus, mit dem die Leistungsmessung beginnt und der Stärke des Anstiegs unterscheiden. Während einige bereits mit einem hohen Ausgangsniveau starten, beginnen andere am Nullpunkt. Probanden des erst genannten Falls weisen öfter Ceilingeffekte auf, da sie das Lernziel mit einer höheren Wahrscheinlichkeit vorzeitig erreichen (vgl. Klauer 2014, S. 12). In einem solchen Fall resultiert ein sogenannter kurvilinearer Verlauf (vgl. Strathmann & Klauer 2010, S. 118).

Strathmann, Klauer und Greisbach (2010) beschreiben vier typische Muster von Verlaufskurven. Das erste Verlaufsmuster wird durch einen linear ansteigenden Verlauf charakterisiert. Es wird ein ständiger Leistungszuwachs, der jedoch auch von mehr oder weniger starken Schwankungen getroffen sein kann, verzeichnet. Dieser Verlaufstyp tritt am häufigsten auf. Das zweite Muster umfasst einen Lernverlauf mit Deckeneffekten. Dies ist unvermeidlich, wenn SuS bereits vor der letzten Lernverlaufsmessung an die obere Grenze des Leistungsspektrums stoßen (vgl. ebd., S. 72). In dem Fall, dass die Zuwächse immer kleiner werden, spricht man auch von negativer Beschleunigung. Neben dem linearen Verlauf erhält der Verlauf dann auch eine quadratische Komponente, die zu dieser sogenannten negativen Beschleunigung führt. Das bedeutet, dass nach starken Fortschritten, der Zuwachs abschwächt und die Leistung sich immer weniger verbessert. Es wird zwar immer weiter geübt, aber ein sichtbarer Erfolg stellt sich nicht mehr ein. In diesem Fall wird auch von „Overlearning“ gesprochen (vgl. Klauer 2006, S. 22).

Der dritte Verlaufstyp zeigt praktisch keine Lernzuwächse. SuS mit diesem Verlaufstyp weisen entweder Verschlechterungen auf, oder sie bleiben konstant auf ihrem Ausgangsniveau. Oft ist dieser Lernverlauf auch von starken Auf-und-Ab-Schwankungen gekennzeichnet, aber ohne an konstantem Lernzuwachs zu gewinnen. Das vierte und letzte Verlaufsmuster beschreibt einen nichtlinearen Verlauf, d.h. einen Verlauf, der durch periodische Schwankungen charakterisiert ist (vgl Strathmann, Klauer & Greisbach 2010, S. 73).

Trotz dieser offensichtlichen Regelmäßigkeiten muss immer mit Tagesschwankungen der SuS gerechnet werden (vgl. Klauer 2006, S. 25). Wurden die Lernverläufe grafisch dargestellt, stellt sich die Frage, wie mit den Daten weiter umgegangen werden kann. Allgemein stehen drei Möglichkeiten zum Umgang mit den Daten zur Verfügung, maintain instruction, change instruction oder increase goal (vgl. Förster et al. 2017, S. 120). Walter (2009) schlägt

für den weiteren Umgang beispielsweise die Auswertung anhand der Drei-Punkte-Regel vor. Dafür muss im Vorhinein eine Ziellinie gesteckt werden. Da in Deutschland keine lange Tradition von Lernverlaufsmessungen vorliegt, liegen folglich bisher auch keine Normen als Orientierungshilfe für durchschnittliche Zuwachsraten vor, sodass der Zielpunkt und der wöchentliche Zuwachs geschätzt werden müssen. Als Anhaltspunkt können aber die Normtabellen von Walter (2008) hinzugezogen werden (vgl. S. 75). Die Ziellinie wird, wie der Lernverlauf, in ein Koordinatensystem eingetragen. Ausgangspunkt der Ziellinie ist der Lernstand des Kindes zu Beginn der Lernverlaufsmessung. Zielpunkt ist beispielsweise die von der Lehrkraft festgesetzte Anzahl von Wörtern, die der Schüler oder die Schülerin bis zum Ende des Jahres pro Minute lesen können soll. Die Auswertung mittels der Drei-Punkte-Regel sieht vor, dass wenn drei Punkte - jeder Punkt steht für das Ergebnis eines Messzeitpunktes - über der Ziellinie liegen, das Ziel höher gesetzt werden sollte (increase goal). Liegen drei Punkte unter der Ziellinie sollten weitere Interventionen einsetzen oder die Fördermaßnahmen geändert werden (change instruction). Dementsprechend meint die dritte Möglichkeit zum Umgang mit den Daten, d.h. maintain instruction, dass die verwendete Fördermaßnahme unverändert bleibt (vgl. Walter 2009, S. 168).

Eine weitere Möglichkeit ist die Auswertung mit der Trendregel, wobei im Gegensatz zur Drei-Punkte-Regel hier „der tatsächliche Lernzuwachs (ermittelt z.B. über regressionsanalytische Verfahren) mit der erwarteten Entwicklung verglichen“ wird. Da in diesem Fall nicht nur die letzten drei Messzeitpunkte berücksichtigt werden, wird meist davon ausgegangen, dass die Trendregeln zu zuverlässigeren Ergebnissen kommen. Die Auswahl der Regeln sollte jedoch individuell je nach Lernverlauf entschieden werden (Förster et al. 2018, S. 120).

3.5 Boden- und Deckeneffekte in den Lernverläufen

Walter (2014) weist daraufhin, dass Verfahren zur Lernverlaufsmessung Boden- und Deckeneffekte möglichst vermeiden sollten und empfiehlt daher Intervallskalen-Niveaus zu nutzen. Denn sollten Boden- und Deckeneffekte auftreten, sinken die Varianz und die Differenzierungskraft der Messergebnisse. Aufgrund des Intervallskalenniveaus soll folglich eine hohe Varianz erreicht werden, indem zwischen den verschiedenen Skalenwerten gleiche Abstände gehalten werden (vgl. S. 171).

Mit der Bezeichnung des Ceiling- oder Deckeneffekts ist gemeint, dass bereits sehr hohe Messwerte sich nicht mehr vergrößern können, da die Messskala des Testverfahrens im oberen Merkmalsbereich begrenzt ist. Die Testergebnisse stoßen somit sozusagen unter die Decke und können die eigentliche Leistung der Probanden nicht mehr wirklichkeitsgetreu darstellen. Der Gegenpart des Deckeneffekts ist der Boden- oder Flooreffekt. Während beim

Deckeneffekt keine zuverlässigen Aussagen über den oberen extremen Merkmalsbereich mehr möglich sind, ist dies beim Bodeneffekt demnach in dem unteren extremen Merkmalsbereich der Fall. Um dies zu vermeiden wird von den in extremen Merkmalsbereichen begrenzten Skalen abgeraten und ein Intervallskalenniveau empfohlen (vgl. Döring & Bortz 2016, S. 738). In der Testwertverteilung erkennt man das Auftreten von Boden- oder Deckeneffekten daran, dass „sich die erzielten Testwerte an einem Ende der Testwertverteilung" häufen, was auch als Testwertstutzung bezeichnet wird (Pospeschill 2010, S. 91).

Nach Lienert und Ratz (1998) resultieren Boden- und Deckeneffekte aus einer „unzweckmäßigen Schwierigkeitsgraduierung" (S. 155). Um Boden- und Deckeneffekte in einem Test zu vermeiden wird daher eine Schwierigkeitsanalyse der Items empfohlen. Verwendete Items dürfen weder zu schwer sein, da sonst Bodeneffekte folgen, noch zu leicht, da im Umkehrschluss sonst Deckeneffekte die Folge sind. Dafür sollte der Schwierigkeitsindex P_i berechnet werden. Moosbrugger und Kelava (2012) beschreiben den Schwierigkeitsindex P_i eines Items i als den „Quotient aus der bei diesem Item tatsächlich erreichten Punktsumme aller n Probanden und der maximal erreichbaren Punktsumme, multipliziert mit 100" (S. 76). Durch die Multiplikation mit dem Faktor 100 liegt der Schwierigkeitsindex bei dichotomer Datenstruktur, wie im hier verwendeten Levumitest, zwischen 0 und 100. D.h. je öfter die Probanden ein Item lösen, desto höher fällt P_i aus. Aufgrund der dichotomen Datenstruktur (0 = falsch oder nicht gelöstes Item; 1 = richtig gelöstes Item) ergibt sich zur Berechnung des Schwierigkeitsindexes P_i eines Items folgende Formel:

$$P_i = \frac{Anzahl\ richtig\ gelöster\ Items}{Anzahl\ aller\ Items} \times 100$$

Ist ein Test zu leicht gestaltet, sodass Deckeneffekte auftreten, entsteht eine rechtsgipfelige Testwertverteilung. Das bedeutet, ein Großteil der SuS löst mehr als die Hälfte der Items korrekt. Umgekehrt folgt eine linksgipfelige Testwertverteilung, wenn die Items zu schwer sind und die meisten SuS weniger als die Hälfte der Items lösen können (vgl. Lienert & Ratz 1998, S. 155). Moosbruger und Kelava (2012) bezeichnen dieses Merkmal in Bezug auf das Auftreten von Deckeneffekten als linksschief, in Bezug auf Bodeneffekte folglich rechtsschief. Die Schiefe der Scoreverteilung kann mit der folgenden Formel berechnet werden:

$$Schiefe\ (x) = \frac{\sum_{v=1}^{n}(x_v - \bar{x})^3}{nSD(x)^3}$$

Fällt das Ergebnis positiv aus, also ist die $\text{Schiefe}\ (x) > 0$, so ist die Verteilung rechtsschief, oder synonym auch linkssteil. Ist die $\text{Schiefe}\ (x) < 0$, d.h. das Ergebnis hat einen negativen

Wert, so ist die Verteilung linksschief, oder rechtssteil (vgl. S. 94). Abgesehen von der Errechnung des Schiefekoeffizienten kann die Schiefe auch an der Beziehung der Lagemaße abgelesen werden. Es bestehen drei Lageregeln, die aufgrund des Verhältnisses von Mittelwert, Modalwert und Median, Aussagen über die Verteilung der gesamten Daten ermöglicht. Die erste Regel ist für den Fall formuliert, dass die drei Lageparameter zusammenfallen. Denn in diesem Fall kann von einer symmetrischen Verteilung gesprochen werden. „Ist der Modalwert kleiner als der Median und dieser wiederum […] kleiner als das arithmetische Mittel“, so liegt nach der zweiten Lageregel eine linkssteile bzw. rechtsschiefe Verteilung vor. Als rechtssteil bzw. linksschief kann eine Verteilung bezeichnet werden, wenn Median und Modalwert größer ausfallen als der Mittelwert (Voß 2015, S. 128).

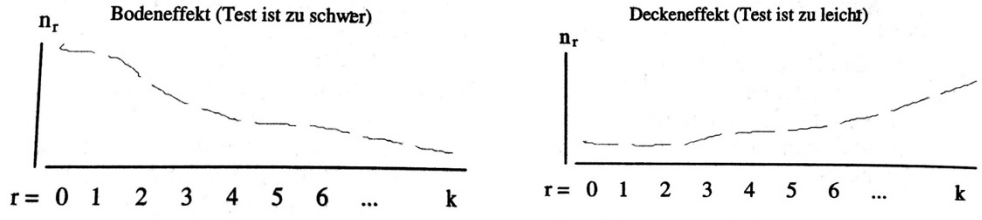


Abbildung 4: *zwei Formen der Scoreverteilungen (entnommen aus Rost 2004, S. 92)*

Abbildung 4 zeigt links eine rechtsschiefe bzw. linkssteile Scoreverteilung, in der ein Bodeneffekt auftritt. Demgegenüber stellt die rechte Grafik eine linksschiefe oder rechtssteile Scoreverteilung dar, die durch einen Deckeneffekt geprägt wird.

3.6 Aktueller Forschungsstand: Lernverlaufsdiagnostik in der Domäne Lesen

In den vorausgegangenen Kapiteln konnten bereits die immense Bedeutung des Lesens, die Komplexität des Leseerwerbsprozesses sowie die Vorteile und der Nutzen der Lernverlaufsdiagnostik verdeutlicht werden. Da die Leistungen deutscher SuS im Bereich Lesen in internationalen Leistungsstudien unterdurchschnittliche Leistungen erzielten, ist eine Förderung der Lesekompetenz über die ganze Schullaufbahn hinweg von fundamentaler Bedeutung. Für eine optimale Förderung ist, wie bereits bewiesen werden konnte, eine gute Diagnostik des aktuellen Leistungsstandes unumgänglich. Aufgrund dessen gibt dieses Kapitel einen Überblick über verschiedene Möglichkeiten zur Erfassung der Lesekompetenz mittels der Lernverlaufsdiagnostik und den aktuellen Forschungsstand zur Lernverlaufsdiagnostik im Lesen.

Diagnostische Lesetests haben im deutschen Sprachraum ihren Ursprung in den 1950er Jahren. Bisher wurden meist standardisierte Lesetests zu unterschiedlichsten Zwecken eingesetzt. So können sie einmal als Diagnoseinventar zur Erfassung der Lernvoraussetzungen

und zur Überprüfung von Fördermaßnahmen eingesetzt werden oder auch zur Ermittlung von Schwierigkeiten und als Leselernverlaufsmessung nützlich sein (vgl. Baumann 2006, S. 869f.). Die Lesefähigkeit deutscher SuS wurde bisher kaum objektiv ermittelt, obwohl Leseschwierigkeiten grundsätzlich mehr fokussiert werden als Rechtschreibschwierigkeiten, was unter anderem durch die hohe Bedeutung der Lesefähigkeit in allen Lebensbereichen begründet wird (vgl. Scheer-Neumann 2006b, S. 552f.). Ein Kritikpunkt an bisher verwendeten diagnostischen Lesetests ist, dass die Qualität des Unterrichts als ein Bestandteil des Leseerwerbs in den Tests nicht weiter beachtet wird. Das bedeutet, sollten Fehlentwicklungen festgestellt werden, werden diese auf individuelle Eigenschaften der SuS zurückgeführt und eben nicht auf unterrichtliche Lerngelegenheiten (vgl. Bredel et al., 2011, S. 168). Nach Scheerer-Neumann liegt die Schwierigkeit der Lesediagnostik außerdem auch darin, dass „sich der Leseprozess im Laufe der Leseentwicklung verändert; die Lesediagnose ist also nur auf dem Hintergrund der Leseentwicklung möglich“ (vgl. Scheer-Neumann 2006b, S. 552f.). Allein anhand dieser zwei genannten Schwierigkeiten der herkömmlichen Lesediagnoseverfahren kann der Nutzen der Lernverlaufsdiagnostik im Lesen deutlich gemacht werden. Denn zum einen zielen die fortlaufenden Lernverlaufsmessungen eben darauf ab, die unterrichtlichen Lerngelegenheiten immer wieder an die Lernprozesse der SuS anzupassen und beziehen mögliche Fehlentwicklungen somit nicht mehr lediglich auf die individuellen Fähigkeiten der SuS zurück. Zum anderen liegt der besondere Vorteil der Lernverlaufsdiagnostik in der Fokussierung des Lernprozesses, sodass nicht nur der aktuelle Leistungsstand diagnostiziert wird, sondern die komplette Leseentwicklung in den Blick genommen wird.

Um die Lesekompetenz eines Schülers oder einer Schülerin mittels der Lernverlaufsdiagnostik zu erfassen, stehen der Lehrkraft grundsätzlich zwei verschiedene Möglichkeiten zur Verfügung. So kann einerseits die Lesefertigkeit, d.h. die Lesegeschwindigkeit und die Leseflüssigkeit erfasst werden und andererseits das Leseverständnis. Da, wie in Kapitel 2.2 schon angesprochen wurde, die Lesefertigkeit als Fundament für das Leseverständnis angesehen werden kann, sollten stets beide Testmöglichkeiten angewendet werden, weil Schwierigkeiten beim Leseverständnis z. B. auch auf Schwierigkeiten in der Leseflüssigkeit zurückgeführt werden können (vgl. Baumann 2006, S. 874). Bezüglich der Erfassung der Lesekompetenz entdeckten Deno, Mirkin und Chiang bereits im Jahr 1982, dass eine hohe Korrelation zwischen der Leseflüssigkeit und dem Leseverständnis vorliegt und identifizierten somit die Leseflüssigkeit als robusten Indikator für die allgemeine Lesekompetenz (vgl. Deno, Mirkin & Chiang 1982, S. 42). Auslassungen, Einfügungen, die falsche oder verzögerte Aussprache von Graphemen oder Silben werden hier als Fehler gewertet, solange sie nicht nachträglich korrigiert werden (vgl. Förster & Souvignier 2011, S. 23). Während Deno (1985) berichtet „we

had not expected to find such a close relationship between reading aloud from text and comprehension scores“ (S. 223), erklärt Sturm (2011), dass der hohe Zusammenhang zwischen der Leseflüssigkeit und dem Textverstehen lediglich während der Grundschulzeit vorliegt und mit steigendem Alter abnimmt (vgl. S. 15). Um die Leseflüssigkeit zu überprüfen, werden den SuS Textabschnitte vorgelegt, die sie in einem Zeitraum von ein bis drei Minuten laut vorlesen müssen (oral reading fluency ORF), wobei die Anzahl der richtig gelesenen Worte pro Minute (WPM) gezählt wird (vgl. Jungjohann et al. 2018a, S. 104). Wayman et al. (2007) konnten in Validitätsuntersuchungen für das Kompetenzmaß der richtig gelesenen Wörter pro Minute Retest-Reliabilitätskoeffizienten zwischen .82 und .97 herausstellen (vgl. S. 85). Eine Schwierigkeit bezüglich des Vorlesens von Textabschnitten liegt allerdings daran, dass es sich als schwierig erweist den SuS Textstellen vorzulegen, die immer denselben Schwierigkeitsgrad haben und nicht dem Kriterium der Homogenität der Testschwierigkeit widersprechen (vgl. Klauer 2011, S. 210). Neben der Erfassung der Leseflüssigkeit mittels ORF-Verfahren listet Walter (2014) zwei weitere Verfahren zur Erfassung der Lesekompetenz auf. Zum einen die maze selection, in der die SuS einen Textabschnitt für einen Zeitraum von einer bis drei Minuten leise für sich lesen. Dieses Testformat testet im Unterschied zu den beiden anderen Verfahren auch das Leseverständnis. Denn an der Stelle jedes siebten Wortes steht eine Klammer mit drei Worten zur Auswahl, jeweils zwei Distraktoren und ein richtiger Begriff (vgl. S. 168). Um die Schwierigkeit zu erhöhen, sind die beiden Distraktoren dem richtigen Begriff entweder semantisch oder phonologisch sehr ähnlich. Das Kompetenzmaß liegt hier in der Anzahl der korrekt ausgewählten Wörter (vgl. Jungjohann et al. 2018a, S. 104). Ein Vorteil dieses Verfahrens ist, dass es auch als Gruppenverfahren eingesetzt werden kann und somit die Praktikabilität und die Zeitökonomie der Tests deutlich erhöht wird (vgl. Walter 2008, S. 77). Weiter listet Walter (2014) die Wort-Identifikation (word identification fluency WIF) auf. In diesem Aufgabentyp müssen die SuS in einem Zeitraum von einer Minute unzusammenhängende Wörter einer Liste laut vorlesen. Auch hier ist das Kompetenzmaß die Anzahl der richtig vorgelesenen Wörter (vgl. S. 168). Die Listen können außerdem auch Silben oder Pseudowörter enthalten. Wie in den ORF-Verfahren wird auch in den WIF-Verfahren die Leseflüssigkeit der SuS erfasst. Im Gegensatz zum Verfahren der maze selection, können das WIF- und ORF-Verfahren bereits zum Schulbeginn verwendet werden, da sie basale Lesekompetenzen erfassen (vgl. Jungjohann et al. 2018a, S. 104).

Eine wesentliche Teilkompetenz bzw. Vorläuferfähigkeit des Leseprozesses ist das phonologische Dekodieren. Um die Beherrschung dieser Teilfertigkeit zu überprüfen, enthalten Lesefertigkeitstests häufig auch Pseudowörter, die wie zu Beginn des Leseerwerbs buchsta-

benweise erlesen werden (vgl. Baumann 2006, S. 875). Denn beherrschen die SuS die Phonem-Graphem-Korrespondenzen, sind sie in der Lage sinnvolle Wörter sowie auch sinnfreie Pseudowörter zu dekodieren (vgl. Schneider 2017, S. 19).

Im Jahr 2010 beschrieb Diehl noch einen Mangel an Diagnoseinstrumenten, die Lernfortschritte im Lesen valide aufzeigen können (vgl. S. 74). Daraufhin entwickelte sie gemeinsam mit Hartke das Inventar zur Erfassung der Lesekompetenz im 1. Schuljahr (IEL-1) (2012), welches als curriculumbasiertes Verfahren über einen Zeitraum von drei Messzeitpunkten den Lernverlauf abbildet. Neben dem IEL-1 stellen auch die bereits erwähnten, von Walter entwickelten Verfahren der Lernfortschrittsdiagnostik Lesen (LDL) (2009) und die Verlaufsdiagnostik sinnerfassendes Lesen (VSL) (2013) gute Möglichkeiten dar, den Lernverlauf auch nach dem ersten Schuljahr zu erfassen.

Inzwischen können auch besonders durch Online-Diagnoseinventare Lernverläufe der SuS optimal erfasst werden. Die Fortschritte bezüglich der digitalen Informations- und Kommunikationstechnologie ermöglichen nun ganz neue Optionen des diagnostischen Handelns in der Schule. Durch die computergestützte Lernverlaufsdiagnostik ergeben sich Vorteile wie eine gute Zeitökonomie sowie eine erleichterte Datenauswertung. Selbstredend muss die Schule dafür aber entsprechend ausgerüstet sein, d.h. über ausreichend PCs und eine gute Internetverbindung verfügen und auch Lehrpersonen müssen eine gewisse Offenheit gegenüber neuen Verfahren aufbringen (vgl. Maier 2014, S. 69f.). Ein Exempel für ein computergestütztes, auf dem Internet basierendes Verfahren zur Lernverlaufsdiagnostik ist quop.de. Entwickelt von Souvignier und Förster wird es bereits seit einigen Jahren in der Schulpraxis genutzt. Neben Tests zur Lesekompetenz für die Jahrgangsstufe 1 bis 6 verfügt quop auch über Testungen für die Fächer Mathematik und Englisch (vgl. Souvignier et al. 2014b, S. 241f.). Auch das Diagnoseinstrument Levumi, welches in der unten vorgestellten Langzeitstudie verwendet wurde, erfasst den Lernverlauf der SuS in allen vier Grundschuljahren über ein internetbasiertes System. Weiteres dazu in Kapitel 5.1.

EMPIRISCHER TEIL

4 Fragestellung und Hypothesen

Zu Beginn des Kapitels 2.2 wurde bereits darauf hingewiesen, dass sich die Lernverläufe der SuS in unterschiedlichem Ausmaß in Abhängigkeit der Lernausgangslage entwickeln. Weisen SuS bereits am Anfang der Förderung ein hohes Lernausgangsniveau auf, entstehen in der Folge häufig Deckeneffekte (vgl. Garbe 2010, S. 16). Da auch Jungjohann et al. (2018a) in ihrem systematischen Review feststellen konnten, dass insbesondere die Testergebnisse älterer SuS, bei denen die *oral reading fluency*, d.h. die Anzahl korrekt gelesener Wörter innerhalb einer Minute, gemessen wurden, Deckeneffekte ergaben (vgl. S. 112), sollen nun auch die Lesetests des Diagnoseinstruments Levumi auf das Auftreten möglicher Decken- und Bodeneffekte hin überprüft werden.

Genauer gesagt wird in der folgenden Studie untersucht, wie stark die Testergebnisse des Diagnoseinstruments Levumi von Boden- und Deckeneffekten geprägt werden. Da die Ausführungen von Jungjohann et al. (2018a) ergaben, dass besonders ältere SuS in Verfahren zur Leseflüssigkeit Deckeneffekte erzielten, wurde für dieses Forschungsvorhaben eine Stichprobe mit Probanden der dritten und vierten Jahrgangsstufe gewählt. Das Diagnoseinventar Levumi enthält jedoch nicht nur Tests zur Überprüfung des Lernverlaufs der Leseflüssigkeit, sondern auch einen Test zum sinnentnehmenden Lesen. Daher wird in dieser Untersuchung die gesamte Lerndomäne des Lesens untersucht, sodass die Leseflüssigkeits- und die sinnentnehmenden Lesetests der Levumi-Plattform auf die Ausprägung enthaltener Boden- und Deckeneffekte untersucht werden. Folglich zielt diese Studie darauf ab folgende Fragestellung zu klären:

> *F1: Wie stark sind Boden- und Deckeneffekte in den Tests der Lernplattform Levumi innerhalb der Lerndomäne „Lesen“ in den Klassenstufen 3/4 einer inklusiven Grundschule ausgeprägt?*

Aus der vorliegenden Fragestellung werden nun zentrale Hypothesen abgeleitet, die mithilfe der folgenden empirischen Untersuchung überprüft werden:

> *H_1: Die Lernplattform Levumi weist innerhalb der Lerndomäne „Lesen“ in den Jahrgangsstufen 3 und 4 einer inklusiven Grundschule Boden- und Deckeneffekte auf.*
>
> *H_2: Die Häufigkeit von Deckeneffekten in den Schülerverläufen nimmt mit fortschreitenden Messzeitpunkten zu, während die Häufigkeit des Auftretens eines Bodeneffekts gleichzeitig abnimmt.*

H_3: Die Lesetests zum Wortlesen und zum sinnentnehmenden Lesen weisen häufiger Deckeneffekte auf, als die Tests zum Silben- und Pseudowörterlesen.

Um mögliche Boden- und Deckeneffekte zu ermitteln und vergleichbare Ergebnisse zu erhalten, wurden alle SuS in jedem der vier Tests sowie zu allen Messzeitpunkten auf der Niveaustufe N4 getestet. Als weitere Komponente der Forschungsfrage wird im Folgenden auch diskutiert, wie mit möglicherweise auftretenden Boden- und Deckeneffekten umgegangen werden sollte und welche Konsequenzen für die Schulpraxis, die Didaktik, aber auch für die weitere Entwicklung des Diagnoseinventars Levumi daraus folgen.

Um Boden- und Deckeneffekte ermitteln zu können, müssen zuerst Grenzwerte bestimmt werden, d.h. es muss entschieden werden, ab wann ein Boden- bzw. Deckeneffekt vorliegt. Weitere Erklärungen dazu folgen in der Ergebnisdarstellung (Kapitel 6).

5 Methodik

Das folgende Kapitel dient dazu den Forschungsablauf zu konkretisieren und zu operationalisieren. Im Weiteren folgt daher zuerst die Darstellung des Forschungsdesigns, d.h. das Erhebungsinstrument, die Online-Plattform Levumi, wird vorgestellt. Im Anschluss folgen die Darlegung der Stichprobe und der zeitlichen und räumlichen Untersuchungsbedingungen sowie eine genaue Beschreibung der Testdurchführung. Abschließend wird außerdem das Vorgehen der Datenauswertung erläutert.

Unter der Operationalisierung wird die Messbarmachung der Fragestellung und der Hypothesen verstanden. Alle in der Fragestellung enthaltenden Merkmale, die zur Beantwortung dieser benötigt werden, müssen präzisiert und identifiziert werden. Die Schwierigkeit innerhalb der Bildungsforschung liegt meist darin, dass die zu messenden Werte nicht direkt beobachtbar sind. Die hier untersuchte Lesekompetenz wird lediglich als ein Konstrukt bezeichnet, da es kein direkt wahrnehmbares Phänomen ist, sondern theoretischen Ursprungs. Um die Lesekompetenz dennoch zu erfassen, müssen messbare und direkt wahrnehmbare Merkmale zur Beschreibung des Konstrukts Lesekompetenz ermittelt werden. Diese messbaren Merkmale werden als Indikatoren bezeichnet (vgl. Sikora 2015, S. 68f.).

Da Deno et al. bereits im Jahr 1982 die Leseflüssigkeit, bzw. die Anzahl der gelesenen Wörter pro Minute, als robusten Indikator für die Lesekompetenz ausmachten (vgl. Kapitel 3.6), wird dieser auch für die Leseflüssigkeitstests innerhalb des Diagnoseinstruments Levumi herangezogen (vgl. S. 42). Hinsichtlich der Tests zum Leseverständnis wird davon ausgegangen, dass das Testverfahren maze selection ebenfalls als robuster Indikator für die Messung des Leseverständnisses angesehen werden kann. Im Unterschied zu den üblicherweise durchgeführten maze selection-Verfahren auf Textebene, finden die sinnentnehmenden Lesetests in Levumi lediglich auf Satzebene statt. Dennoch müssen die SuS ebenfalls aus einer Auswahl von vier Antwortmöglichkeiten das richtige Wort erkennen.

Während sich die Operationalisierung der Lesetests leicht gestaltet, da das Diagnoseinstrument diese vorgibt, gestaltet sich die Operationalisierung der Boden- und Deckeneffekte schwieriger. Um diese zu operationalisieren, müssen Grenzwerte gesetzt werden. Es muss bestimmt werden, wie viele Aufgaben jedes einzelnen Tests gelöst werden müssen, um von einem Decken- bzw. Bodeneffekt sprechen zu können.

5.1 Darstellung des Forschungsdesigns

Durch die Wahl des Forschungsdesigns ergeben sich bereits bestimmte Vorgaben für die Aussagen der folgenden Forschung (vgl. Knigge 2015, S. 57f.). Das folgende Forschungsdesign ist durch die Bezeichnung des Ex-post-facto-Designs erklärbar. Denn im Unterschied zu experimentellen und quasi-experimentellen Forschungsdesigns, wird innerhalb des Ex-post-facto-Designs kein aktives Treatment vorgenommen. Ein Nachteil von Ex-post-facto-Designs ist ihre Schwäche hinsichtlich Kausalitätsaussagen, ein Vorteil hingegen, dass es aufgrund des nicht vorhandenen Treatments nah an den tatsächlichen Begebenheiten ist. Mittels des Ex-post-facto-Designs können Quer- und Längsschnittuntersuchungen durchgeführt werden. Die folgende Studie ist als Langzeitstudie, oder genauer gesagt als Panel-Studie angelegt, da der Lernverlauf der SuS über ein Schuljahr hinweg betrachtet wird. Denn nach Auffassung von Knigge (2015) zeichnet sich eine Panel-Studie[4] dadurch aus, dass die gleichen Probanden über einen längeren Zeitraum zu mehrmaligen Messzeitpunkten befragt werden, um Daten zu Veränderungen eines bestimmten Merkmals zu erhalten. Ein Problem der Panel-Studien ist die Panel-Mortalität, d.h. einige Probanden fallen aus unterschiedlichen Gründen aus der Stichprobe heraus (vgl. S. 64f.). Die Panel-Mortalität trat auch in dieser Studie auf, sodass die Stichprobengröße über die vier Messzeitpunkte hinweg schwankt.

5.1.1 Aufbau der Online-Lernplattform Levumi

Wie im Theorieteil bereits mehrfach angeklungen, wurde die Ermittlung von Decken- und Bodeneffekten in der Lernverlaufsmessung mit dem Erhebungsinstrument Levumi durchgeführt. Die Lernplattform stellt wie von Diehl (2011) gefordert ein „whole-in-one-Paket“ dar (vgl. Gebhardt & Jungjohann, 2018). Denn durch die Verbindung eines Diagnoseinstruments mit entsprechenden Fördermaterialien, wie durch das Forschungsprojekt Levumi bereitgestellt, kann eine optimale Förderung erfolgen (vgl. Diehl 2011, S. 171). Die Diagnoseplattform Levumi wird im Folgenden näher dargestellt.

Trotz des in den letzten Jahren stark gestiegenen Interesses an der Verlaufsdiagnostik, ist das Angebot vorhandener Diagnoseinstrumente bisher eher gering (vgl. Voß & Gebhardt 2017, S. 95), wodurch der Online-Lernplattform Levumi (**Le**rn-**V**erla**u**fs-**M**on**i**toring) eine besondere Bedeutung zukommt. Als ein gemeinsames und offenes Forschungsprojekt wurde die Lernplattform www.levumi.de von den Wissenschaftlern Markus Gebhardt (TU Dortmund), Kirsten Diehl (Europa-Universität Flensburg) und Andreas Mühling (Christian-Alberts-

[4] Anderen Autoren nach kann die Panel-Studie auch umfangreicher sein, hier lediglich auf die Erklärung der Bezeichnung nach Knigge (2015) bezogen.

Universität zu Kiel) entwickelt (vgl. Gebhardt et al. 2016a, S. 1) und stellt somit eine Zusammenarbeit von empirischer Bildungsforschung, fachdidaktischer Forschung und Informatik dar (vgl. Mühling et al. 2017, S. 557). Levumi wurde für alle 16 Bundesländer Deutschlands entwickelt und orientiert sich an den Lernzielen dieser. Besonders werden SuS mit Lernschwierigkeiten und Verhaltensauffälligkeiten fokussiert. Die Plattform zeichnet sich durch eine sehr einfache Bedienung und eine leicht verständliche Darstellung aus, die durch ein frei erhältliches Lehrerhandbuch (Gebhardt et al. 2016a), Tutorials und ein Handbuch zu Förderansätzen im Lesen (Jungjohann et al. 2017) zusätzlich unterstützt wird (vgl. Jungjohann, DeVries, Gebhardt & Mühling 2018b, S. 3). Die Tests können außerdem beispielsweise durch die Anpassung der Schriftgröße im Sinne des „Universal Designs“ personalisiert werden (vgl. Mühling et al. 2017, S. 559). Zentrale Ziele des Diagnoseinstruments sind

> (1) to offer teachers a practical CBM tool for inclusive class-rooms, (2) to improve research on CBM and the acceptance of CBM tools by teachers, and (3) to use the collected data for evaluating supporting materials for research and development in teaching and learning (Jungjohann et al. 2018b, S. 3).

Im Gegensatz zu traditionellen Paper-Pencil-Tests sowie gegenüber dem bereits seit längerer Zeit in der Praxis angewandten internetbasierten Diagnoseinstruments quop, hat das ebenfalls auf dem Internet basierenden Diagnoseinstrument Levumi den Vorteil, dass die Benutzung komplett kostenfrei ermöglicht wird. Die einzige Voraussetzung, unter der Bedingung des Vorhandenseins von nutzbaren Computern, ist ein Internetzugang; empfohlen wird der Zugang über den Webbrowser *Mozilla Firefox*. Da die Daten auf den Servern der beteiligten Universitäten gespeichert werden, sind keine Installationen von Software-Programmen o.ä. auf den schuleigenen PCs notwendig. Ein weiterer Vorteil liegt in der immensen Aufwandserleichterung für die Lehrkräfte, da ihnen die Auswertung und Darstellung der Daten abgenommen wird. Nichtsdestotrotz ist auch eine Testdurchführung als Paper-Pencil-Variante möglich. In diesem Fall müssen die Ergebnisse jedoch händisch ins Computersystem eingetragen werden. Denn die Auswertung mit Levumi hat den Vorteil, dass eine Klassen- und eine Schüleransicht dargestellt wird. Das bedeutet, dass ein individueller Leistungsvergleich, aber auch ein sozialer Vergleich ermöglicht werden (vgl. Gebhardt 2016b, S. 447). Nach Maier (2010) sind *short-cycle*- und *medium-cycle*-Rückmeldungen besonders erfolgsversprechend im Hinblick auf Verbesserungen der Schülerleistungen. Von short-cycle-Rückmeldungen wird gesprochen, wenn das Feedback unmittelbar nach den Testungen erfolgt (vgl. S. 302). In der Levumi-Plattform ist dies der Fall, da das Maskottchen, der Drache Levumi, direkt nach der Testung den SuS eine Reaktion auf Verbesserungen zeigt (vgl. Jungjohann et al. 2018b, S. 3). Die Entwickler der Onlineplattform bauen das Instrument stetig weiter aus und können auch beim Auftreten von Bedienungsfehlern oder anderen auftretenden

Problemen in der Praxis kontaktiert werden und stehen stets zur Unterstützung bereit (vgl. Gebhardt et al. 2016b, S. 447). Da das Forschungsprojekt Levumi bisher noch in seinen Anfängen steht, sind noch einige Fragen bezüglich der Umsetzung oder der Gütekriterien offen. Für die Silbentests konnte aber bereits eine gute Test-Retestreliabilität sowie die Testfairness für SuS mit und ohne Förderbedarf bestätigt werden. Außerdem wurde auch festgestellt, dass sich der Test eignet um den Lernprozess über mehrere Messzeitpunkte hinweg zu testen (vgl. Jungjohann et al. 2018b, S. 6). Während die Levumi-Plattform lediglich mit Tests zur Leseflüssigkeit startete, umfasst sie inzwischen Tests zum sinnentnehmenden Lesen, zur Rechtschreibung sowie zum Wortschatz. Darüber hinaus enthält Levumi auch bereits drei verschiedene Tests zum Unterrichtsfach Mathematik. Weiterhin werden zeitnah auch Tests zum Schülerverhalten sowie Tests für die Sekundarstufe implementiert (vgl. Jungjohann et al. 2018b, S. 7). Im anschließenden Subkapitel wird sich auf die Lerndomäne des Lesens fokussiert, da in der empirischen Studie ausschließlich die Tests zur Lesekompetenz verwendet wurden.

5.1.2 Die Domäne „Lesen“ in der Plattform

Die Online-Lernplattform beinhaltet innerhalb der Domäne „Lesen“ derzeitig Lernverlaufstests zur Leseflüssigkeit (Silben-, Wörter- und Pseudowörterlesen) und einen Test zur sinnentnehmenden Lesekompetenz (vgl. Jungjohann & Gebhardt, 2018, S.165). Zur Unterstützung der Lehrkräfte wurde außerdem ein Lehrerhandbuch entwickelt, welches konkrete Fördermaßnahmen zur Leseförderung mit Levumi enthält (vgl. Jungjohann et al. 2017, S. 1). Die Leseflüssigkeitstests wurden mittels des oral reading fluency-Verfahrens durchgeführt. Das bedeutet, die SuS müssen in einem Zeitraum von einer Minute so viele Items, welche auf dem Computerbildschirm nacheinander erscheinen, laut vorlesen, wie sie innerhalb dieser Zeitbegrenzung schaffen. Der Testleiter entscheidet dann darüber, ob das Item korrekt (= 1) oder nicht korrekt (= 0) vorgelesen wurde. Die Leseflüssigkeitstests laufen dementsprechend nur mittels eines Raters statt. Im Gegensatz dazu ist der Test zum Leseverständnis im Sinne der Testform maze selection entwickelt worden. Da die SuS diese selbstständig am PC bearbeiten können, kann dieser Test auch als Gruppentest durchgeführt werden (vgl. Jungjohann et al. 2018a, S. 104). Während der Testentwicklung der Lesetests wurde sich am Lehrplan für Grundschulen sowie an dem Kieler Leseaufbau orientiert (vgl. Jungjohann et al. 2017, S. 1).

Der Kieler Leseaufbau (KLA) hat seinen Ursprung eigentlich in der Legasthenie-Therapie. Um Levumi im Unterricht zu etablieren, muss der KLA nicht in die Unterrichtspraxis übernom-

men werden. Die Levumi-Entwickler empfehlen jedoch den KLA als Strukturierungshilfe einzusetzen (vgl. Gebhardt 2016 et al. S. 448). Die Wirksamkeit des Förderprogramms auf die Leseleistung konnte bereits bewiesen werden (vgl. Galuschka & Körne 2015, S. 483). Der KLA zeichnet sich durch sein kleinschrittiges und strukturiertes Vorgehen aus. Denn das Förderprogramm gibt in 14 Schwierigkeitsstufen die Abfolge der Buchstabeneinführung vor, sodass Schwierigkeiten möglichst vermieden werden (vgl. Dummer-Smoch & Hackethal 2007, S. 14).

Wie im CBM vorgesehen erhält jeder Schüler und jede Schülerin bei Levumi seinen eigenen Test, d.h. eine eigene Zufallsstichprobe an Items, sodass das Testverfahren immer gleich schwere Paralleltest entwickelt. Um auch schwache SuS oder SuS mit sonderpädagogischem Förderbedarf optimal zu testen, kann zwischen unterschiedlichen Schwierigkeitsniveaus gewählt werden (vgl. Gebhardt et al. 2016b, S. 447f.). Die Schwierigkeitsstufen in Levumi (N0 bis N4) sind in Anlehnung an die 14 Schwierigkeitsstufen des KLA entwickelt worden. Um auch mit anderen Lehrwerken kompatibel zu sein, wurde die Reihenfolge der Buchstabeneinführung jedoch leicht verändert (vgl. Jungjohann et al. 2017, S. 3). Durch die Orientierung der Niveaustufen am KLA wird auf die Erfüllung des Testkriteriums der Homogenität der Testschwierigkeit abgezielt. Tabelle 1 gibt einen Überblick über die Stufen des KLA im Vergleich zu den Niveaustufen in Levumi.

Tabelle 1

Aufbau des KLA und der darauf aufbauenden Schwierigkeitsstufen in der Plattform Levumi (in Anlehnung an Jungjohann et al. 2017, S. 4)

eingeführte Buchstaben im KLA	Niveaustufen in Levumi	eingeführte Buchstaben in Levumi
Vorstufe	**N0**	m, l – a, e, i, o, u
a, e, i, o, u, au, ei		
1-2	**N1**	m, r, s, n, f, l – a, e, i, o, u
m, r, s, n, f, l		
4-5	**N2a**	h, w, s, p, t, d – en, er, el – a, e, i, o, u, ei, au
ch, w, z, p, t, k		
6-7	**N2b**	ch, k, b, sch, g – a, e, i, o, u
b, d, g, eu, sch, el		
8-10	**N3 (a/b)**	j, v, ß, sp, st (ohne tz, ck), z, qu, x,y – eu, ä, ö, ü
j, v, ß, ä, ö, ü, qu, x, y		– a, e, i, o, u
11-14	**N4**	alle Buchstaben

In Levumi wird die Schwierigkeit nicht nur durch die Wahl der Buchstaben hergestellt, sondern auch durch die Wahl der Tests. Die Schwierigkeit steigt von den Silbentests über die

Wörtertests hin zu den Pseudowörtertests (vgl. Jungjohann et al. 2017, S. 4). Während die Tests zum Silbenlesen das lautierte Lesen überprüfen, testen die Wörtertests das lexikalische Lesen. Die Pseudowörter werden aus den Silben neu und sinnfrei zusammengesetzt. Mit diesen Tests kann das nichtlexikalische Lesen, d.h. die Dekodierfähigkeit der SuS gemessen werden (vgl. Gebhardt et al. 2016, S. 448). Die drei Tests, die die Kompetenz der Leseflüssigkeit testen, nutzen für die Messung den von Deno et al. (1982) identifizierten robusten Indikator (vgl. Jungjohann et al. 2018b, S. 4). Abbildung 5 verdeutlicht die Struktur der Lerndomäne Lesen und gibt ebenfalls einen Überblick über die verschiedenen Niveaustufen der einzelnen Tests.

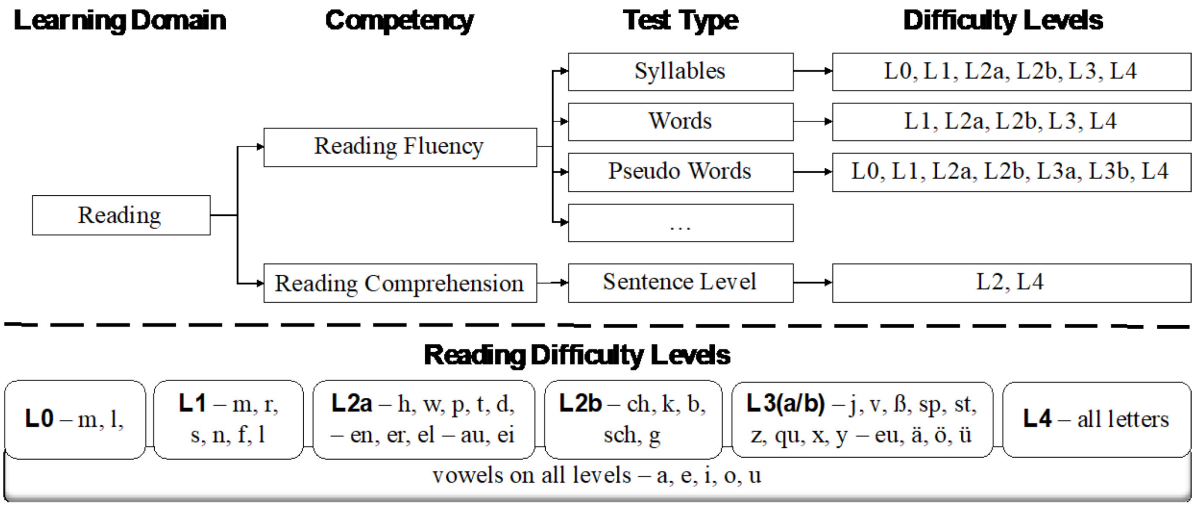


Abbildung 5: *Test Structure of the Learning Domain Reading in the Levumi platform (vgl. Jungjohann et al. 2018b, S. 4)*

Bezüglich der Leseverständnistests werden im Unterschied zum üblichen maze selection-Verfahren keine ganzen Texte, sondern lediglich Sätze vorgegeben, in denen die SuS das fehlende Wort aus einer Auswahl von vier Wörtern ergänzen müssen. Zwischen den einzelnen Sätzen sowie zwischen den Wörtern der Wortlesetests besteht kein Sinnzusammenhang. Der Leseprozess wird daher nicht durch den zu erwartenden Inhalt erleichtert (vgl. Braun 2010, S. 178). In allen Testbereichen erhalten alle SuS zum ersten Testzeitpunkt die gleiche Stichprobe und die gleiche Reihenfolge der Items, erst während des zweiten Messzeitpunktes erhält jeder eine zufällige Ziehung.

5.2 Darlegung der Stichprobe, des Untersuchungssettings und -zeitraums

Die Langzeitstudie wurde in einer inklusiven Grundschule in einem ländlichen Gebiet in Nordrhein-Westfalen über einen Zeitraum von einem Schuljahr durchgeführt. Während des Schuljahrs 2017/2018 wurde die Grundschule von drei Master-Studentinnen der Technischen Uni-

versität Dortmund zu vier Messzeitpunkten besucht. Die Messzeitpunkte wurden stets so gewählt, dass sie jeweils ein paar Wochen vor den Ferien stattfanden, um sicher zu gehen, dass Ferieneffekte vermieden werden. Denn die Festlegung der Messzeitpunkte in die Wochen vor den Ferien wird auch von dem Administratorenteam empfohlen (vgl. Gebhardt 2016a, S. 5). Der erste Messzeitpunkt fand an drei Tagen der 42. Kalenderwoche des Jahres 2017, d.h. Mitte Oktober vor den Herbstferien, statt. Jeder Messzeitpunkt fand über jeweils drei Tage und je über die erste bis zur fünften Schulstunde statt. Messzeitpunkt 2 wurde dementsprechend zwei Wochen vor den Weihnachtsferien durchgeführt. Nach dem Jahreswechsel fand der dritte Messzeitpunkt im März 2018 eine Woche vor den Osterferien statt. Der letzte Messzeitpunkt fand aufgrund der Vielzahl an anderweitigen Aktivitäten am Ende des Schuljahres bereits in der ersten Juniwoche statt. Zwischen den Messzeitpunkten fand kein universitär beeinflusstes oder kontrolliertes Treatment statt.

Um die Studie möglichst störungsfrei und zeitökonomisch durchführen zu können, stellte die Schule für die Testungen einen Computerraum zur Verfügung. Der Computerraum umfasste bestenfalls 14 funktionierende PCs.

Die Forschungsstichprobe umfasst insgesamt N= 87 SuS der dritten und vierten Jahrgangsstufe. An der Studie teilgenommen haben zwei dritte und zwei vierte Jahrgangsstufen, sodass die SuS zwischen neun und elf Jahren alt waren. Während die Jahrgangsstufe drei mit n= 47 (54,02%) SuS vertreten ist, nahmen n= 40 (45,98%) SuS der Jahrgangsstufe vier teil. Von diesen SuS waren n= 41 (47,13%) weiblich und n= 46 (52,87%) männlich. Insgesamt betrug der Anteil von SuS mit Migrationshintergrund 12,64 %, d.h. 11 SuS. N = 4 SuS wurde ein Förderschwerpunkt zugeschrieben, davon n= 2 der Förderschwerpunkt Deutsch[5], n= 1 der Förderschwerpunkt Lernen und ebenfalls n= 1 der Förderschwerpunkt Körperliche und motorische Entwicklung.

Aufgrund von Messfehlern wurden die Messergebnisse zweier Schüler aus der Wertung genommen. Ein weiterer Messfehler könnte darin bestehen, dass während des vierten Messzeitpunktes aufgrund eines Systemfehlers einige SuS einer dritten Klasse mit der gleichen Wortreihenfolge getestet wurden wie im ersten Messzeitpunkt. Da aufgrund des langen dazwischen liegenden Zeitraums ein Übungseffekt ausgeschlossen werden konnte, wurden diese Daten auch weiterhin in der Auswertung berücksichtigt.

[5] Es ist bekannt, dass der Förderschwerpunkt Deutsch offiziell nicht existiert, im Rahmen von Levumi wurde sich jedoch dafür entschieden diesen angeben zu können, um den Lehrkräften eine Möglichkeit zu geben zu signalisieren, dass diese SuS Schwierigkeiten beim Erwerb der L2-Deutsch aufweisen, oder aktuell erst die Sprache erwerben.

Die Anzahl der einzelnen Testergebnisse schwankt teilweise jedoch sehr stark. Dies liegt entweder daran, dass einzelne SuS erkrankt waren und nicht an allen Testterminen teilnehmen konnten, oder daran, dass aufgrund technischer Probleme und aufgrund von Schwierigkeiten resultierend aus der schlechten Internetverbindung nicht alle Testergebnisse gespeichert wurden.

5.3 Durchführung

Die Messzeitpunkte liefen alle nach demselben Schema ab. Die Klassen wurden je nach Klassengröße in zwei oder drei Schulstunden nacheinander abgearbeitet. Begonnen wurde jedes Mal mit dem Test zur sinnentnehmenden Lesefähigkeit. Dafür wurde zuerst die gesamte Klasse mittels alphabetischer Reihenfolge in zwei Lerngruppen, bzw. drei bei größeren Klassen, eingeteilt. In jeder Klasse wurde mit der Gruppe des hinteren Alphabets begonnen. Denn da der Computerraum maximal 14 PCs bereithält, konnte nicht direkt die gesamte Klasse getestet werden. Zur Zeitersparnis wurden die SuS bereits im Vorhinein von den Testleitern ins System eingeloggt, sodass den SuS nur noch der richtige Platz zugewiesen werden musste. Die sinnentnehmenden Lesetests fanden auf Niveaustufe N4 statt, sodass die SuS innerhalb von 7 Minuten maximal 61 Items lösen konnten. Um die Durchführungsobjektivität hochzuhalten, wurden die wörtlichen Durchführungsinstruktionen des Levumi-Programms (Stand 2017, s. Anhang 1) zur Einführung in die Tests befolgt. Um sicher zu gehen, dass die SuS den Testablauf verstanden haben, wurde die Beispielaufgabe gemeinsam im Plenum besprochen. Danach führten die SuS die Lesetests selbstständig an den schuleigenen Computern durch. Nachdem der sinnentnehmende Lesetest durchgeführt wurde, wurden die SuS wieder zurück in ihre Klasse gebracht und die nächste Lerngruppe mit in den Computerraum genommen. Auch diese wurden mit den Durchführungshinweisen über den Ablauf des Tests aufgeklärt und lösten ihre Tests eigenständig. Von dieser zweiten Gruppe wurden jeweils sechs SuS, d.h. die ersten sechs der alphabetischen Klassenliste, im Computerraum behalten, während die anderen SuS begleitet zurück in die Klasse gehen konnten. Die sechs SuS, die im Computerraum verblieben, wurden direkt im Anschluss in ihrer Leseflüssigkeit getestet. Es wurden die 1-Minute-Lesetests im Wortlesen (61 Items), im Silbenlesen (134 Items) und im Pseudowortlesen durchgeführt (189 Items). Die sechs SuS wurden so zugeteilt, dass jede der drei Testleiterinnen jeweils zwei SuS nacheinander testen musste. Die Verteilung der SuS auf die Studentinnen blieb nach Möglichkeit über alle Messzeitpunkte hinweg so erhalten. In dem Fall, dass SuS aus Krankheitsgründen oder ähnlichem fehlten, wurde aus zeitökonomischen Gründen die Zuteilung leicht verändert. Auch die Reihenfolge der drei

Tests zur Leseflüssigkeit wurde für jede 6er-Gruppe anders bestimmt, änderte sich aber zwischen den Messzeitpunkten nicht. Der Schüler bzw. die Schülerin, der/die nicht getestet wurde, wurde aufgefordert sich auf die andere Seite der Studentin zu setzen und sich leise zu verhalten. Da während der ersten Testung die Itemreihenfolge in allen Einzeltests gleich ist, ergibt sich für den zuschauenden Schüler an dieser Stelle möglicherweise ein geringer Vorteil. Dies ist aber, wie bereits erklärt, lediglich während des ersten Messzeitpunktes der Fall. Aufgrund der zu langsamen Internetverbindung der Schule, wurden die Tests zur Leseflüssigkeit an den Laptops der Studentinnen über einen mobilen W-LAN-Router durchgeführt.

5.4 Vorgehen der empirischen Auswertung

Bezüglich der Auswertung von empirischen Daten können verschiedene Wege eingeschlagen werden. Eine Möglichkeit der Auswertung ist die Deskriptivstatistik. Die erhobenen Daten werden zusammengefasst, verdichtet und systematisch geordnet, um einen Überblick über die gemessenen Verhältnisse zu bekommen. Die beschreibende Statistik zeichnet sich durch ihre konkrete Sprache sowie durch tabellarische und weitere grafische Darstellungen, z. B. zu Häufigkeitsverteilungen, aus. Häufig berechnete Werte sind Minimal-/Maximalwerte, Mittelwerte und Standardabweichungen (vgl. Orthmann Bless 2015, S. 106f.).

So wird auch die Auswertung der in dieser Studie erhaltenen Daten als Deskriptivstatistik vorgenommen. Dafür wurde mit den beiden Softwareprogrammen Microsoft EXCEL und IBM SPSS gearbeitet. Die Datenaufbereitung, wie beispielsweise die Umwandlung der Reaktionszeiten in die dichotome Datenstruktur (0 = falsch gelöstes Item, 1 = richtig gelöstes Item) wurde mittels des Programms SPSS durchgeführt, während die weitere deskriptive statistische Analyse mit EXCEL durchgeführt wurde. Die Auswertung der Levumi-Daten kann als Speed- oder als Powertest durchgeführt werden. Diese Studie verfolgt den Ansatz des Speedtestes, sodass aufgrund der zeitlichen Beschränkung nicht gelöste Items als falsch gelöste Items gewertet werden. Speedtests sind so gestaltet, dass die Probanden ohne die zeitliche Beschränkung alle Items lösen könnten, d.h. die Testergebnisse werden durch die beschränkte Bearbeitungszeit differenziert (vgl. Preckel & Brüll 2008, S.58).

Im Folgenden werden zuerst die Forschungsergebnisse der einzelnen Tests nacheinander dargestellt. Da die Größe der Stichprobe zwischen den Messzeitpunkten, z.B. aufgrund von krankheitsbedingtem Fehlen der SuS, aber auch zwischen den einzelnen Testformaten innerhalb eines Messzeitpunktes aufgrund von technischen Speicherschwierigkeiten

schwankt, also die Panel-Mortalität hier zutrifft, wird die Stichprobengröße bzw. die Teilstichprobengröße immer auch in Prozentwerten angegeben, damit eine größere Vergleichbarkeit der Ergebnisse erreicht wird.

Die Ergebnisse der einzelnen SuS werden als Personen- oder synonym auch als Summenscore angegeben. Dies bezeichnet die Anzahl der richtig gelösten Items innerhalb eines Tests. Häufig wird der Summenscore auch „Personenfähigkeit“ genannt. Aufgrund unterschiedlicher Auffassungen bezüglich dieser Bezeichnung wird im Folgenden hauptsächlich der Terminus des Summenscores verwendet (vgl. Rost 2004, S. 92).

Die Ergebnisdarstellung beginnt im Folgenden mit den Ergebnissen der Tests zum Leseverständnis. Alle Ergebnisse werden jahrgangsstufenweise vorgestellt, so werden immer erst die Ergebnisse der dritten Jahrgangsstufe und daran anschließend die Ergebnisse der Jahrgangsstufe 4 dargelegt. Die Summenscores aller SuS der einzelnen Messzeitpunkt werden mittels der grafischen Darstellung eines Boxplots präsentiert. Denn auf diese Weise wird es möglich, Minimal- und Maximalwerte, den Median sowie die Streuungsmaße der Quartilsabstände, der Spannweite und der Standardabweichung übersichtlich darzustellen. Um die Testwerte adäquat darzustellen, werden neben dem Median auch die beiden weiteren Lagemaße, d.h. der Mittel- und Modalwert, ermittelt. Da das Verteilungsmaß der Schiefe Auskunft über den Schwierigkeitsgrad der Items ermöglicht und der Schwierigkeitsgrad wiederum Aussagen über das Auftreten von Boden- und Deckeneffekten ermöglicht, werden im Folgenden auch der Schiefekoeffizient sowie, wenn es sich anbietet, auch die Itemschwierigkeit berechnet. Die Itemschwierigkeit bzw. der Schwierigkeitsindex P_i wird in dieser Arbeit lediglich als ein weiteres Merkmal zur Betrachtung der Unterschiede zwischen SuS, die Boden- oder Deckeneffekte aufweisen, angesehen und nicht weitergehend vertieft (vgl. Pospeschill 2010, S. 90). Zur Berechnung dessen wird sich auf die Erklärungen von Moosbrugger und Kelava (2012) bezogen, dabei ist wichtig zu erwähnen, dass das hier gewählte Vorgehen aufgrund der hohen Komplexität nicht dem der Itemanalyse der IRT entspricht (vgl. S. 77).

Das Auftreten von Decken- und Bodeneffekten ist abhängig von den gesetzten Kriterien, ab wann diese Effekte vorliegen. Da diese Kriterien nicht einheitlich für alle Testformate gesetzt werden können, werden zur Bestimmung des Boden- oder Deckeneffekts in dem Test zum Leseverständnis andere Kriterien herangezogen als zur Bestimmung in den drei Leseflüssigkeitstests.

Um die Ausprägung der Boden- und Deckeneffekte in den einzelnen Testformaten darzustellen, werden Punktdiagramme erstellt. Um die Abbildungen nicht zu unübersichtlich zu gestalten, aber auch nicht zu viele Abbildungen aufzuführen, zeigt das Punktdiagramm zu jedem

Test lediglich die Ergebnisse bzw. die Boden- und Deckeneffekte des ersten und vierten Messzeitpunktes. Denn diese beiden Messungen sind von besonderer Relevanz, da sie die Entwicklung über die gesamte Langzeitstudie am besten verbildlichen können. Darstellungen zu jedem einzelnen Messzeitpunkt befinden sich im Anhang (ab S. XVIII).

6 Darstellung der Forschungsergebnisse

Testergebnisse des sinnentnehmenden Lesens – N4

Begonnen wird mit der Darstellung der Testergebnisse des sinnentnehmenden Lesetests. Zur Wiederholung: Die Tests zur sinnentnehmenden Lesefähigkeit enthalten insgesamt 61 Items. Die Testergebnisse des Leseverständnisses in der dritten Jahrgangsstufe schwanken stark. Die Extremwerte spannen über eine Weite von minimal 5 richtig ausgewählten Wörtern bis hin zu der maximalen Höchstpunktzahl von 61 richtig ausgewählten Wörtern. Damit erreichen die Testergebnisse zu Testzeitpunkt 1 und 4 eine maximale Spannweite von 55 Items. Wird dies in Bezug zu der maximal lösbaren Itemanzahl von 61 gesetzt, wird offensichtlich, dass die Ergebnisse der dritten Jahrgangsstufe im sinnentnehmenden Lesetest nahezu über das gesamte Spektrum der angebotenen Items schwankt (s. Abbildung 6 oder Anhang, S. XVIII). Wie ebenfalls Abbildung 6 entnommen werden kann, nehmen die oberen Quartilsabstände stetig ab während die Medianwerte (Md) steigen, was bedeutet, dass immer mehr Teilnehmer immer höhere Summenscores erreichen.

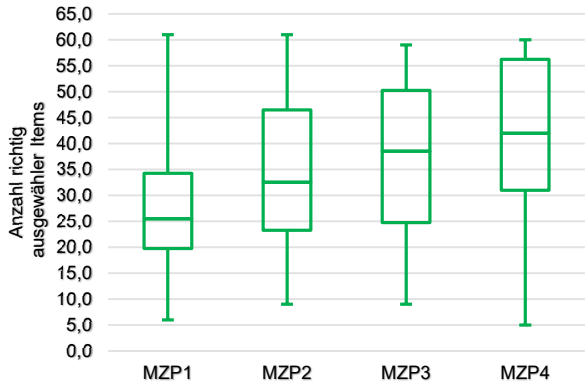


Abbildung 6: *Boxplot – sinnentnehmendes Lesen (Jgst. 3)*

Bei Betrachtung der Anzahl der durchschnittlich gelösten Items, d.h. der Mittelwert (M) (s. Tabelle 2, dritte Zeile) oder auch der Medianwerte (s. Abbildung 6) ist ein deutlicher Zuwachs der Anzahl der korrekt gelesenen Items zu vernehmen. Da auch zum letzten Messzeitpunkt hin ein starker Anstieg der Mittelwerte stattgefunden hat und auch die Standardabweichung (SD) relativ konstant bleibt und somit keine abnehmende Varianz wahrzunehmen ist, lassen sich Boden- und Deckeneffekte erst einmal ausschließen. Wird jedoch der Modalwert in den Blick genommen, d.h. „der häufigste Testwert in der Verteilung“ (Moosbrugger & Kelava 2012, S. 93), wird offensichtlich, dass dieser während des dritten und des vierten Messzeitpunktes bei einer insgesamt erreichbaren Anzahl von 61 mit Mo= 59 (MZP3) und Mo= 58 (MZP4) sehr hoch liegt. So erreichen während des dritten Messzeitpunktes n= 5 Probanden

(≈ 11,36%) den Summenscore von 59, dies entspricht ca. 96,72% der möglichen richtig lesbaren Items. Auch der starke Anstieg des Modalwerts von dem zweiten Messzeitpunkt (Mo= 29) hin zu Messzeitpunkt 3 (Mo= 59) ist hier auffällig.

Tabelle 2

Überblick über die Ergebnisse des sinnentnehmenden Lesetests (Jgst. 3)

	MZP1	MZP2	MZP3	MZP4	
N	44	46	44	42	44
M	27,41	34,21	36,93	41,17	34,93
Md	25,5	32,5	38,5	42	33
Mo	29	29	59	58	59
SD	13,01	14,47	15,37	14,64	13,87

Um weitere Vermutungen über auftretende Boden- und Deckeneffekte treffen zu können, ist es hilfreich sich die Scoreverteilung anzuschauen (s. Abbildung 7). Nach den Lageregeln zur Beschreibung der Symmetrien bedeutet eine linksschiefe Verteilung, dass mehr Werte größer als der Mittelwert sind und der Medianwert demnach höher liegt als der Mittelwert; eine rechtsschiefe Verteilung meint im Umkehrschluss, dass der Mittelwert größer ist als der Median (vgl. Moosbrugger & Kelava 2012, S. 94). Da der Schiefekoeffizient in dieser Testung mit (x) = 0,1 > 0 vorliegt, kann von einer leicht rechtsschiefen Verteilung ausgegangen werden. Dies beweisen auch die Testergebnisse in denen der Median zu Messzeitpunkt 1 (Md= 25,5) unter dem Mittelwert liegt (M≈ 27,41), gleiches gilt für den zweiten Messzeitpunkt (Md= 32,5; M≈ 34,22). Erst zum dritten Messzeitpunkt hin hat der Median (Md= 38,5) den Mittelwert (M≈ 36,93) knapp überholt und hält sich auch während der letzten Testung konstant über dem Mittelwert (Md= 42; M≈ 41,17). Insgesamt liegt der Mittelwert mit M≈ 34,85 knapp über dem Median von Md= 33, dies gibt aber noch keinen Aufschluss über mögliche Boden- und Deckeneffekte.

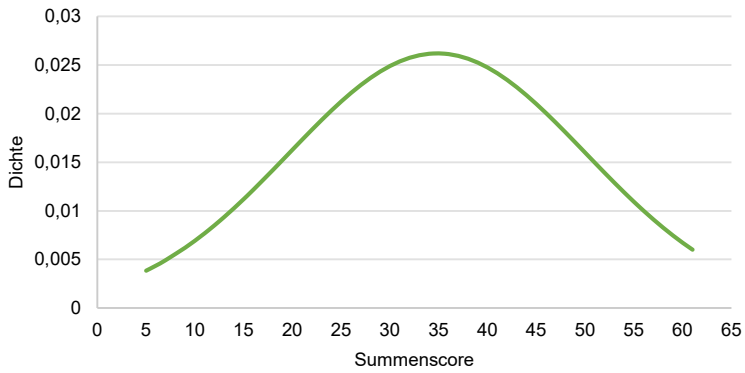


Abbildung 7: *Normalverteilung - sinnentnehmendes Lesen (Jgst. 3)*

Abbildung 7 zeigt nun die Scoreverteilung aller vier Messzeitpunkte zusammen. Dadurch, dass die Kurve rechts leicht abgeschnitten ist und somit zensierte Daten bzw. eine Testwertstutzung vorliegen, wird ein Kennzeichen von Deckeneffekten erfüllt.

Um Boden- und Deckeneffekte letztendlich ermitteln zu können bzw. wie die Fragestellung es vorsieht, Aussagen darüber treffen zu können, wie stark diese ausgeprägt sind, müssen bestimmte Grenzwerte gesetzt werden. Ab wann kann davon gesprochen werden, dass ein Bodeneffekt vorliegt, ab wann kann von einem Deckeneffekt gesprochen werden? Diese Grenzsetzung ist immer abhängig von bestimmten Kriterien. Da für die Levumi-Tests bisher keine Normwerte und auch keine bezüglich der Testzeit und des Testformates vergleichbaren Normwerte anderer Tests vorliegen, werden die Grenzen aufgrund üblicherweise angeführten Begründungen aufgeführt.

Innerhalb des sinnentnehmenden Lesetests gestaltet sich die Aufgabe der Festlegung der Grenze des Bodeneffekts als relativ einfach. Denn da die SuS in diesem Testformat die richtige Antwort aus einer Auswahl von vier Antwortmöglichkeiten herausfinden müssen, liegt allein die Wahrscheinlichkeit die richtige Antwort aus Zufall zu treffen bereits bei 25%. Daher wird für dieses Testformat die Grenze des Bodeneffekts ebenfalls bei 25% gesetzt. Das bedeutet, dass Summenscores von 0 bis zu 15 richtig gelösten Items als Bodeneffekt gezählt werden. Von Deckeneffekten wird eigentlich nur dann gesprochen, wenn die Probanden 100% erreichen, da aber, wie Klauer (2006) erklärt immer mit Tagesschwankungen der Schülerleistungen zu rechnen ist (vgl. S. 25), wird die obere Grenze bei 95% gesetzt, sodass die auftretenden Schwankungen in den Schülerleistungen das Kriterium für den Deckeneffekt darstellen. Ein Deckeneffekt liegt in dem sinnentnehmenden Lesetest also ab einem Summenscore von etwa 58 richtig gelösten Items vor. Als Begründung für dieses Kriterium können die Leistungsschwankungen einer Schülerin angeführt werden: Denn während des ersten Messzeitpunktes erreicht sie die Höchstpunktzahl von 61 richtig gelesenen Items und somit 100%, in den weiteren Testungen erreicht sie jedoch lediglich Summenscores von 58 (MZP2), 59 (MZP3) und 60 (MZP4), ihre Ergebnisse schwanken demnach immer zwischen der 95%- und der 100%-Grenze.

Abbildung 8 stellt nun die Summenscores des ersten Messzeitpunktes sowie des vierten Messzeitpunktes unter Berücksichtigung der genannten Grenzwerte dar. Im Anhang (S. XVIII) befinden sich weitere Darstellungen zu der Ausprägung der Boden- und Deckeneffekte für die einzelnen Messzeitpunkte.

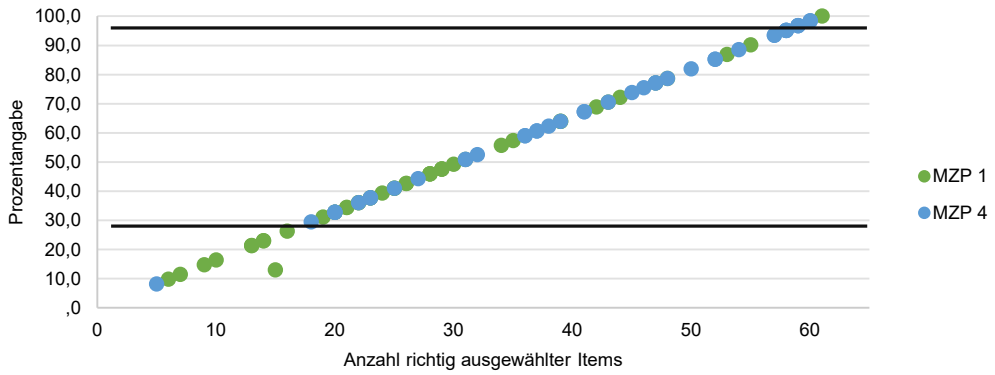


Abbildung 8: *Visualisierung der Ausprägung der Boden- und Deckeneffekte – sinnentnehmendes Lesen (Jgst. 3)*

Schon bei bloßer Betrachtung von Abbildung 8 wird sichtbar, dass die Summenscores des ersten Messzeitpunktes zum Großteil unter der Bodeneffektgrenze liegen, während über der Deckeneffektgrenze deutlich mehr Ergebnisse des vierten Messzeitpunktes liegen. Genauer gesagt liegen die Testergebnisse von n= 9 Probanden (≈ 20,45%) während des ersten Messzeitpunktes unter der Grenze des Bodeneffekts. Durch den bereits beschrieben Zuwachs nimmt die Anzahl der SuS, die unter der Bodeneffektgrenze liegen über die Messzeitpunkte hinweg stetig ab (MZP 2: n= 5 (≈ 10,87%); MZP 3: n= 4 (≈ 9,09%)), sodass zum Zeitpunkt der vierten Testung hingegen nur noch ein einziger Schüler (≈ 2,38%) unter dieser Grenze liegt. Im Hinblick auf die Deckeneffekte lässt sich ein anderes Ergebnis verzeichnen. Während in der ersten Testung lediglich eine Schülerin (≈ 2,27%) aufgrund des Erreichens des maximalen Wertes in den Bereich des Deckeneffektes fällt, sind es während der vierten Testung bereits n= 9 (≈ 21,43%) aller Probanden. Auch diese Veränderung kündigt sich während des Verlaufs der Testungen bereits an. Denn so weisen während der zweiten Testung bereits zwei Probanden (≈ 4,35%) und während der dritten Testung n= 5 (≈ 11,36%) Probanden einen Deckeneffekt auf. Insgesamt weisen demnach in allen Testungen folglich 17 Summenscores (≈ 9,66%) Deckeneffekte und 19 Testergebnisse (≈ 10,8%) Bodeneffekte auf. Würde die Deckeneffektgrenze bei 90% anstatt 95% gesteckt werden, würden diese Ergebnisse noch deutlicher ausfallen. Denn allein beim dritten Messzeitpunkt erreichen ca. 20,45% (n= 9) aller SuS mehr als 90% (≈ 55 richtig gelöste Items) und liegen somit knapp unter der für diese Studie gesetzten Grenze von 95%. Für alle Messzeitpunkte insgesamt betrachtet ergibt sich ein Bild von n= 27 Probanden (≈ 15,34%), die mehr als 90% der Items richtig auswählen konnten.

Wird nun geschaut wie sich die Lernverläufe der SuS entwickeln und ob SuS, die zum ersten Messzeitpunkt von Bodeneffekten betroffen waren, diese auch im weiteren Verlauf der Studie aufweisen, fällt bereits vom ersten zum zweiten Messzeitpunkt nicht nur auf, dass die Anzahl

der Summenscores, die unter dieser Grenze liegen, sich zu Messzeitpunkt 2 beinahe halbiert hat (MZP 1: 9 SuS (≈ 20,45%); MZP 2: 5 SuS (≈ 10,87%)), sondern dass vier der fünf SuS der zweiten Testung auch bereits zu Messzeitpunkt 1 unter dieser Grenze lagen. Gleiches gilt für die Veränderungen hin zur dritten Testung, da wiederum drei SuS die schon zur ersten Testung unter der Grenze lagen wieder dort zu finden sind. Aufgrund dieser Ergebnisse stellt sich die Frage, ob die SuS tatsächlich aufgrund ihrer Leistung bzw. aufgrund des Rateniveaus konstant unter diese Grenze fallen oder ob motivationale Probleme dahinter stecken und die SuS möglicherweise „einfach nur weiterklicken“ ohne wirklich zu lesen. Werden diese SuS genauer betrachtet, fällt auf, dass einem dieser Schüler, der konstant vom ersten bis zum dritten Messzeitpunkt unter der Grenze des Bodeneffekts liegt, der Förderschwerpunkt Deutsch[6] zu geschrieben wurde. Von einem anfänglichen Summenscore von 6 nahm seine Leistung über die Messzeitpunkte jedoch stetig zu (MZP 2: 9, MZP 3: 12), sodass er im vierten Messzeitpunkt einen Summenscore von 20 erreichen konnte und somit 32,79% der möglichen Items korrekt löste.

Bezüglich der anderen SuS, die konstant unter diese Grenzen waren, konnten die weiteren Schülerdaten nichts auffälliges beweisen. Aber auch diese SuS zeigten während der folgenden Testungen immer höhere Summenscores. Interessant ist außerdem das Ergebnis eines Schülers, der während der ersten zwei Messzeitpunkte Summenscores von 47 und 45 erreichte, während der dritten und vierten Testung allerdings nur noch einen Summenscore von 9 und 5 erreicht hat. Da über seine Schülerbiographie auch nichts auffälliges bekannt ist, müssen in diesem Fall motivationale Schwierigkeiten vermutet werden.

Neben dem genannten Schüler, dem der Förderschwerpunkt Deutsch zugeschrieben wurde und der folglich auch einen Migrationshintergrund hat, weisen in dieser Jahrgangsstufe sieben weitere SuS einen Migrationshintergrund auf. Drei dieser SuS weisen überdurchschnittliche Leistungen auf, die weiteren durchschnittliche bis unterdurchschnittliche. Das genaue Verhältnis dieser SuS sollte jedoch genauer untersucht werden.

Hinsichtlich der Deckeneffekte lässt sich im Vergleich zu den Bodeneffekten ein umgekehrtes Bild verzeichnen. Mit den Messzeitpunkten nimmt auch die Anzahl von Summenscores, die über der oberen Grenze liegen, stark zu. SuS, die während der ersten Testungen bereits im Bereich des Deckeneffekts gefallen sind, erzielen auch im Weiteren konstant hohe Werte.

[6] Hier sei noch mal darauf hin gewiesen, dass es bekannt ist, dass der Förderschwerpunkt Deutsch offiziell nicht existiert, im Rahmen des Diagnoseinventars Levumi wurde sich jedoch dafür entschieden diesen angeben zu können, um den Lehrkräften eine Möglichkeit zu geben zu signalisieren, dass diese SuS Schwierigkeiten beim Erwerb der L2-Deutsch aufweisen, oder aktuell erst die Sprache erwerben.

Bei Betrachtung der Ergebnisse des Leseverständnistests der Jahrgangsstufe 4 zeichnet sich im Hinblick auf die Ausprägung der Deckeneffekte ein noch deutlicheres Bild ab. Dies wird bereits in Abbildung 9 deutlich. Denn besonders der Messzeitpunkt 4 zeigt, durch die sehr kleine Box, den abnehmenden oberen Quartilsabstand sowie den sehr hoch gelegenen Medianwert, die stark abgenommene Varianz der Ergebnisse. Der komplette Kasten des Boxplots von Messzeitpunkt 4 liegt wie der Abbildung 9 entnommen werden kann über der der Linie von 55 richtig ausgewählten Items. Das bedeutet, dass 75% der Probanden zu Messzeitpunkt 4 einen Summenscore von mehr als 55 erreichen.

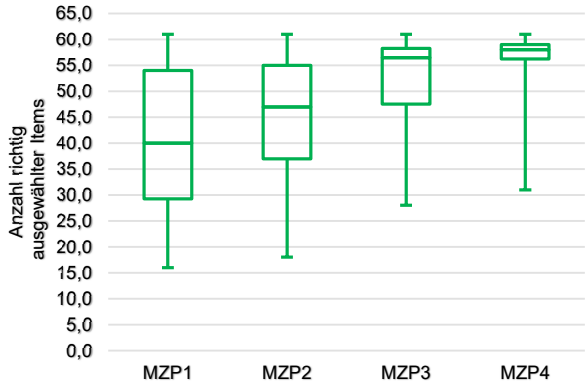


Abbildung 9: *Boxplot - sinnentnehmendes Lesen (Jgst. 4)*

Die Spannweite der Ergebnisse ist im Vergleich zur dritten Jahrgangsstufe bereits immens gesunken, so reicht sie in der Jahrgangsstufe 4 lediglich von 16 korrekt ausgewählten Wörtern bis hin zur maximalen Anzahl von 61 korrekt gelösten Items und liegt somit nur noch bei 45 Items (s. Abbildung 9 oder im Anhang, S. XIX). Auch Tabelle 3 zeigt aufgrund der über die vier Messzeitpunkte stets abnehmende Standardabweichung die sinkende Varianz der Ergebnisse und erfüllt somit ein Kennzeichen für das Auftreten von einem Deckeneffekt.

Tabelle 3

Überblick über die Ergebnisse des sinnentnehmenden Lesetests (Jgst. 4)

	MZP1	MZP2	MZP3	MZP4	
N	38	33	40	40	37,75
M	39,89	45,61	52,1	54,63	48,28
Md	40	47	56,5	58	54
Mo	24	60	59	57	59
SD	13,49	10,68	8,9	7,82	11,9

Ähnlich wie in der Jahrgangsstufe 3 nehmen Mittelwert und Modalwert über die Messzeitpunkte hinweg stark zu und liegen von Beginn an über denen der Jahrgangsstufe 3 (ausgenommen der Modalwert vom ersten und vierten Messzeitpunkt). Über den Verlauf der Messungen nähern sich die Testergebnisse immer weiter der maximalen Itemanzahl von 61 an.

Im Unterschied zur Jahrgangsstufe 3 erreicht der Modalwert hier bereits während des zweiten Messzeitpunktes einen sich an den maximalen Wert annäherndes Ergebnis von Mo = 60 und steigt hier folglich von der ersten zur zweiten Testung bereits stark an und steigt nicht wie in der Jahrgangsstufe 3 erst von dem zweiten zum dritten Messzeitpunkt stark an.

Auch die Testergebnisse der Jahrgangsstufe 4 werden nach ihrer Scoreverteilung und somit nach möglichen Anzeichen für das Auftreten von Boden- und Deckeneffekten untersucht. Abbildung 10 verdeutlicht eindeutig, dass die Kurve der Verteilung der Testergebnisse linksschief bzw. rechtsgipflig ist, da „unverhältnismäßig viele Pbn alle oder fast alle Aufgaben lösen (ceiling effect eines zu leichten Tests)“ (Lienert & Raatz 1998, S. 156). Auch die Berechnung der Schiefe zeigt mit einem negativen Wert von (x) = - 0,85, dass eine linksschiefe, d. h. rechtssteile Verteilung vorliegt. Demnach liegt der Median (Md= 54), entsprechend der Lageregeln, in diesen vier Testungen insgesamt auch höher als der Mittelwert (M≈ 48,28). Anders als in der Jahrgangsstufe 3 liegt der Median in den Testergebnissen der Jahrgangsstufe 4 bereits während allen vier Testzeitpunkten über dem Mittelwert (MZP 1: Md= 40, M≈ 39,89; MZP 2: Md= 47, M≈ 45,61; MZP 3: Md= 56,5, M≈ 52,1; MZP 4: Md= 8, M≈ 54,63). Auch der Graph der Abbildung 10 lässt auf das Auftreten von Deckeneffekten schließen. Denn da dieser in Höhe des maximalen Summenscores von 61 abgeschnitten ist und somit zensierte Daten vorliegen, wird ein Merkmal des Deckeneffekts erfüllt.

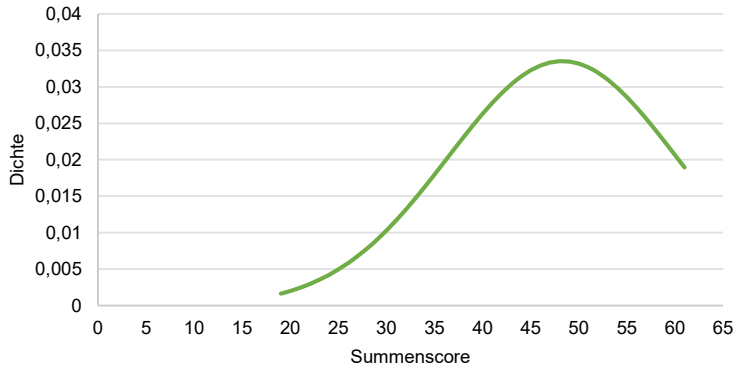


Abbildung 10: *Normalverteilung - sinnentnehmendes Lesen (Jgst. 4)*

Wie bereits Abbildung 8 die Ausprägung der Boden- und Deckeneffekte innerhalb des sinnentnehmenden Lesetests in der dritten Jahrgangsstufe zeigte, stellt Abbildung 11 diese dementsprechend für die Jahrgangsstufe 4 vor. Bereits auf den ersten Blick ist klar, dass kein Testergebnis unter die Grenze des Bodeneffekts von 25% fällt. Die Deckeneffekte scheinen zuerst einmal ähnlich wie in der Jahrgangsstufe 3 auszufallen. Werden jedoch die einzelnen Werte angeschaut (s. Anhang, ab S. XVIII) weisen schon während des ersten Messzeitpunktes n= 6 SuS (≈ 15,79%) einen Deckeneffekt auf. Während der zweiten Testphase treten

lediglich bei n= 5 Probanden (≈ 15,15%) ein Deckeneffekt auf, zu dieser Testung lag jedoch mit 33 Teilnehmern auch die kleinste Stichprobengröße vor. Zum Testzeitpunkt 3 liegen n= 14 SuS (≈ 35%) über der Deckeneffektgrenze, bei Testzeitpunkt 4 sind es dann bereits mehr als die Hälfte der Teilnehmer (n= 22 (≈ 55%)). Das bedeutet, dass während allen vier Messungen insgesamt 47 von 151 Testergebnissen (≈ 31,13%) Deckeneffekte aufweisen – eindeutig mehr als in der Jahrgangsstufe 3 (≈ 9,66%).

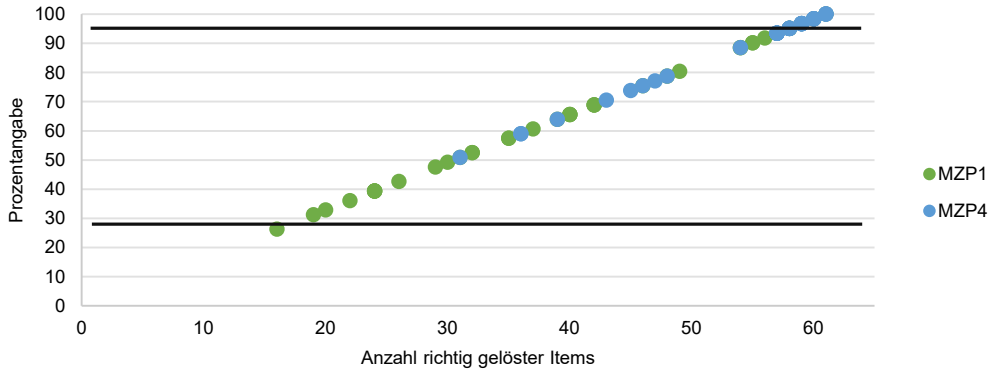


Abbildung 11: *Visualisierung der Ausprägung der Boden- und Deckeneffekte – sinnentnehmendes Lesen (Jgst. 4)*

Wird auch hier geschaut, ob einige SuS nur knapp unter der Grenze des Deckeneffekts liegen, wie es Abbildung 11 zufolge scheint, also einen Summenscore von 55 (≈ 90%) oder mehr erreichen, erhält man zu Messzeitpunkt 1 und 2 nur jeweils n= 9 Probanden (≈ 23,68% bzw. ≈ 27,27%) auf die dies zutrifft. Im dritten Messzeitpunkt liegen n= 24 SuS (=60%) knapp und der hier eigentlich gesetzten Grenze von 95%, in der vierten Testung sind es n= 30 Teilnehmer (≈ 75%), die einen Summenscore von 55 oder mehr erreichen, wie es auch bereits dem Boxplot des vierten Messzeitpunktes (s. Abbildung 9) zu entnehmen ist. Daraus ergibt sich, dass während der gesamten Langzeitstudie 72 von 151 Testergebnissen (≈ 47,68%) knapp unter der Deckeneffektgrenze liegen.

In der Jahrgangsstufe 4 wurde drei SuS ein Förderschwerpunkt zugeschrieben. Ein Schüler mit diagnostiziertem Förderschwerpunkt Körperliche und motorische Entwicklung weist konstant überdurchschnittliche Ergebnisse auf. Sein Summenscore schwankt während den Messzeitpunkten zwischen 58 und 59 und liegt somit auch im Bereich des Deckeneffekts. Ein Schüler mit diagnostiziertem Förderschwerpunkt Lernen weist, abgesehen von Messzeitpunkt 3, durchgängig unterdurchschnittliche Leistungen auf. Meist liegt er etwa eine Standardabweichung unter dem Mittelwert. Einer dritten Schülerin wurde von den Lehrkräften der Schule der Förderschwerpunkt Deutsch zugeschrieben. Diese weist während allen Testzeit-

punkten stark unterdurchschnittliche Summenscores auf, von einem anfänglichen Summenscore von 22 steigen ihre Leistungen aber konstant an bis zu einem Summenscore von 36 zu Messzeitpunkt 4. Mit dieser Schülerin weisen auch drei weitere SuS einen Migrationshintergrund auf. Die Ergebnisse dieser Schülergruppe fallen mit mindestens einer Standardabweichung anfangs auch unterdurchschnittlich aus, steigen aber stetig an, sodass die Summenscores von zwei dieser drei SuS während des letzten Messzeitpunktes ebenfalls im Bereich des Deckeneffekts liegen. Aufgrund der geringen Anzahl dieser SuS können diese Daten aber nicht als repräsentativ angesehen werden.

Insgesamt kann für die sinnentnehmenden Lesetests der Levumi-Plattform festgehalten werden, dass die Deckeneffekte eindeutig stärker ausgeprägt sind als die Bodeneffekte, welche trotz der hoch gesetzten Grenze in der Jahrgangsstufe 4 gar nicht mehr auftreten. Dass die Deckeneffekte stärker ausgeprägt sind, gilt im besonderen Ausmaß für die vierte Jahrgangsstufe, in welcher die Ergebnisse des vierten Messzeitpunktes von 55% der Probanden durch den aufgetretenen Deckeneffekt begrenzt werden. Der Umgang mit diesen Ergebnissen wird im anschließenden Diskussionskapitel weiter vertieft.

Eine Ursache für das Auftreten von Decken- und Bodeneffekten kann darin liegen, dass der Test zu leicht oder zu schwer für die Schülergruppe ist. Daher ist es auch von Bedeutung sich die Itemschwierigkeiten anzuschauen. Da die Plattform Levumi im Sinne des Speedtests konstruiert wurde und die Homogenität der Testschwierigkeit im Sinne des CBM ein wichtiges Testkriterium darstellt, sollten eigentlich alle Items die gleiche Schwierigkeit aufweisen. In der Speedtest-Variante zählt jedoch nicht die Itemschwierigkeit, sondern die Menge der Items. Denn Speedtests differenzieren die Ergebnisse nicht nach dem Schwierigkeitsniveau der gelösten Items, sondern nach der Anzahl der gelösten Items. Aufgrund der starken Ausprägung des Deckeneffekts wird hier dennoch die Itemschwierigkeit in den Blick genommen. So lässt sich im Hinblick auf die Itemschwierigkeit innerhalb des sinnentnehmenden Lesetests kein Unterschied zwischen den falsch gelösten Items von SuS, deren Lernverläufe Bodeneffekte aufweisen und den falsch gelösten Items von SuS, deren Lernverläufe von Deckeneffekten geprägt werden, entdecken. Auffällig ist hingegen, dass über alle vier Messzeitpunkte hinweg und in beiden Jahrgangsstufen einige Items auffallend seltener gelöst werden als die anderen. Obwohl der Schwierigkeitsindex P_i unter 33 liegen muss, um als schweres Item zu gelten, was zumindest in der Jahrgangsstufe 4 nicht zutrifft, sind diese Items aufgrund ihres, im Vergleich zu den anderen Items, stets schlechteren Abschneidens zu betrachten. Neben den Items „weder“ und „über“, ist insbesondere das Item „In“ (Aufgabenstellung des Items „In“: *Aus/In/Durch/Im* Schloss wohnt ein Geist.) auffällig seltener gelöst worden und ist in beiden

Jahrgangsstufen sowie über alle Messzeitpunkte hinweg immer das Item, welches am seltensten gelöst wurde (vgl. Tabelle 4). Bei Betrachten von Tabelle 4 ist es jedoch von Bedeutung, dass die Itemschwierigkeiten zu Messzeitpunkt 1 anders zu bewerten sind. Denn während in den Messzeitpunkten 2, 3 und 4 für alle SuS eine andere Reihenfolge der Items hergestellt wurde, sieht das Diagnoseinventar Levumi vor, dass alle SuS zum ersten Messzeitpunkt dieselbe Reihenfolge an Items erhalten. Das bedeutet, dass sich beispielsweise das bessere Abschneiden des Items „In“ zum ersten Messzeitpunkt in der Jahrgangsstufe 3 damit erklären lässt, dass dieses Item das sechste Item des Tests ist und somit noch von allen SuS bearbeitet wurde. Items, die in der Reihenfolge des ersten Tests weiter hinten liegen, wie beispielsweise das Item „über“ fallen zum ersten Messzeitpunkt dementsprechend schlechter aus, da nicht alle SuS in dem begrenzten Zeitraum bis zu diesem Item gelangen. Ein Überblick zum Vergleich aller Itemschwierigkeiten befindet sich im Anhang (S. XXX).

Tabelle 4

Vergleich auffällig schwerer ausfallender Items („In“, „über“, „wenn“, „weder“) in den sinnentnehmenden Lesetests mit einem konstant leicht ausfallenden Item („gut“ bzw. *„wenn“)*

Jahrgangsstufe 3				
Items	MZP 1	MZP 2	MZP 3	MZP 4
	P_i	P_i	P_i	P_i
In	36,36	17,39	15,9	38,1
über	18,18	34,78	40,9	40,48
wenn	15,91	36,96	72,73	47,52
gut	72,73	71,74	61,36	78,57
Jahrgangsstufe 4				
Items	MZP 1	MZP 2	MZP 3	MZP 4
	P_i	P_i	P_i	P_i
In	34,21	33,33	40	62,5
über	34,21	60,61	70	77,5
weder	55,26	72,72	72,5	87,5
wohne	65,79	87,88	97,5	92,5

Die durchschnittliche Itemschwierigkeit innerhalb der sinnentnehmenden Lesetests in dem Diagnoseinventar Levumi schwankt zwischen den beiden Jahrgangsstufen. In der Jahrgangsstufe 3 lag der durchschnittliche Schwierigkeitsindex bei P_i = 57,26, in der Jahrgangsstufe 4 hingegen bei und P_i = 78,78. Nach Döring und Bortz (2016) wird ein Schwierigkeitsindex zwischen P_i = 20 und P_i = 80 empfohlen, sodass die durchschnittliche Itemschwierigkeit der Jahrgangsstufe 3 in den Optimalbereich fällt. Auch der Schwierigkeitsindex für die Stufe 4 liegt im erwünschten Bereich, liegt jedoch schon sehr nah am Bereich der extrem leichten Items.

Testergebnisse des Leseflüssigkeitstests zum Silbenlesen - N4

Die Festlegung der Grenzen in dem Levumi-Test zum Leseverständnis gestaltet sich aufgrund der logischen Schlussfolgerung bezüglich der Grenze des Bodeneffekts relativ einfach. Diese Grenze kann jedoch nicht für die Tests der Leseflüssigkeit übernommen werden, sodass hier insbesondere ein neuer Grenzwert für den Bodeneffekt ermittelt werden muss. Der Grenzwert des Deckeneffekts bei 95% kann mit derselben Begründung wie bei den sinnentnehmenden Lesetests auch hier übernommen werden. Das bedeutet folglich, dass die SuS mindestens 127 Items richtig vorlesen müssen, um einen Deckeneffekt auszulösen, da der Silbentest insgesamt 134 Items umfasst. Von Bodeneffekten spricht man im ursprünglichen Sinn, wenn der Test ungeeignet ist, um schlechtere Schülerleistungen zu erfassen, d. h., dass die SuS nicht in der Lage sind, Items zu lösen. In dieser Arbeit wird die Grenze jedoch bei 5% angelegt, da dies lediglich einer Anzahl von sieben Items anspricht und die Schülerleistungen im Silbentest ebenfalls häufig um eine Anzahl von sieben Items schwanken.

Die Silbentests umfassen 134 Items. Die Spannweite der Ergebnisse der Jahrgangsstufe 3 streckt sich über alle Messzeitpunkte hinweg über eine Weite von 12 (MZP 1) bis hin zu 76 (MZP 4) richtig gelesenen Silben (s. Abbildung 12). Das bedeutet einerseits, dass keiner der SuS auch nur annähernd die maximale Anzahl von 134 Items erreicht. Andererseits zeigt dies die große Streuung der Ergebnisse, die auch durch die über alle Messzeitpunkte hohe Standardabweichung bewiesen wird (s. Tabelle 5). Die höchste Spannweite der Summenscores liegt mit einem Unterschied von 54 Items zu Messzeitpunkt 2 und 4 vor.

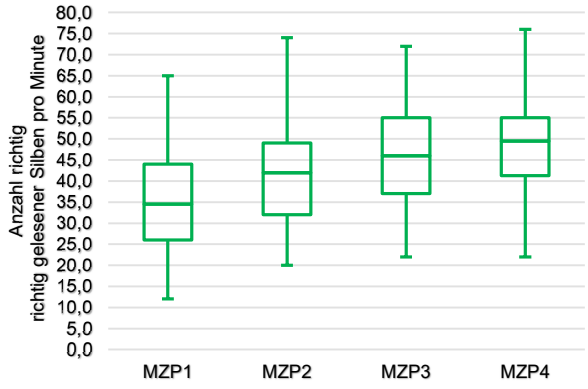


Abbildung 12: *Boxplot - Silben lesen (Jgst. 3)*

Die durchschnittliche Anzahl richtig vorgelesener Silben steigt von M= 35 in Messzeitpunkt 1 hin zu M= 48 richtig gelesenen Silben in Messzeitpunkt 4 konstant an. Der Modalwert

schwankt hingegen während der Langzeitstudie zwischen Mo= 34 und Mo= 50, sodass der Modalwert lediglich zu Messzeitpunkt 4 über dem Mittelwert liegt (s. Tabelle 5).

Tabelle 5

Überblick über die Ergebnisse des Silbentests (Jgst. 3)

	MZP1	MZP2	MZP3	MZP4	M
N	46	45	44	45	45
M	35	43	46	48	43
Md	35	42	46	50	43
Mo	34	42	37	50	40
SD	11,76	13,96	13,28	11,69	12,67

Beim Vergleich von Tabelle 5 mit Abbildung 12 fällt auf, dass zu Messzeitpunkt 1 und zu Messzeitpunkt 2 Mittelwert, Modalwert und Medianwert sehr nah bei einander liegen. Während die Werte im Messzeitpunkt 1 bei M= 35, Mo= 34 und Md= 35 liegen, liegen sie im zweiten Messzeitpunkt leicht gesteigert bei M= 43, Mo= 42 und Md= 42. Nach Moosbrugger und Kelava (2012) spricht man in dem Fall, dass die drei Lageparameter zusammenfallen, von einer symmetrischen Verteilung (vgl. S. 93). Das bedeutet, dass die Wahrscheinlichkeit für ein Testergebnis, das über den drei Maßen liegt, genauso hoch ist, wie die Wahrscheinlichkeit, dass es unter diesen liegt. Für die Daten dieser Erhebung bedeutet dies, dass in etwa genauso viele SuS besser als die drei Maße abschneiden, wie Testergebnisse, die schlechter als diese ausfallen. Da der Schiefekoeffizient jedoch bei (x) ≈ 0,16 liegt, bedeutet dies, dass eine leicht rechtsschiefe Verteilung vorliegt (s. Abbildung 13). Im Gegensatz zu den Ergebnissen der sinnentnehmenden Lesetests wird die Kurve hier nicht begrenzt, da der Test eigentlich 134 Items erfasst, ist insbesondere nach rechts noch viel Spielraum. Dieses Ergebnis spricht gegen das Auftreten von Boden- und Deckeneffekten.

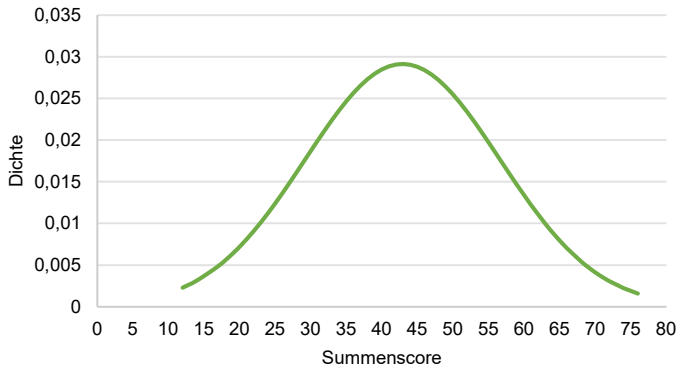


Abbildung 13: *Normalverteilung – Silben lesen (Jgst. 3)*

Auch bei Betrachtung der festgesetzten Grenzen der Boden- und Deckeneffekte in Abbildung 14 lässt sich, wie bereits kurz erwähnt, feststellen, dass während des ersten und vierten Messzeitpunktes kein Proband unter oder über die festgesetzten Grenzen fällt.

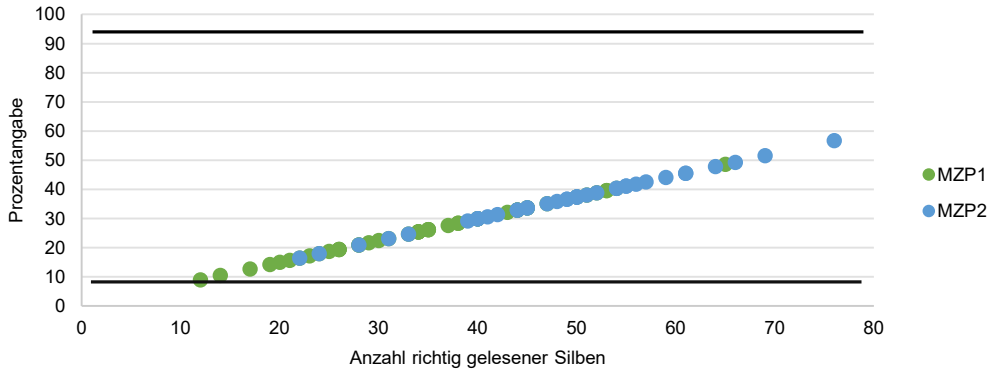


Abbildung 14: *Visualisierung der Ausprägung der Boden- und Deckeneffekte – Silben lesen (Jgst. 3)*

Der höchste Testwert, der in allen vier Testungen erreicht wurde, beträgt 76, dies entspricht ca. 56,72% der gesamten Itemanzahl. Der geringste Testwert liegt bei 12 richtig erlesenen Silben in einem Zeitraum von einer Minute (≈ 8,96%), sodass folglich während allen Testzeitpunkten kein Bodeneffekt auftritt. Würde die Bodeneffektgrenze jedoch bei 10% angesetzt sein, würde das genannte Ergebnis, als einziges Ergebnis der gesamten Langzeitstudie, unter diese Grenze fallen. Auch wenn die Grenze des Bodeneffekts auf 15% angehoben werden würde, wird sich an diesen Ergebnissen nicht viel ändern. Denn lediglich 7 von insgesamt 180 (≈ 3,89%) ermittelten Summenscores im Silbenlesen liegen unter der Grenze von 15%, dies entspricht etwa 20 richtig gelesenen Silben innerhalb einer Minute. Der Großteil dieser Ergebnisse lässt sich im ersten Messzeitpunkt auffinden, während der dritten und vierten Testungen liegen alle Summenscores über der 15%-Grenze.

In den Ergebnissen des Silbenlesens in der dritten Jahrgangsstufe lassen sich demnach keine Boden- und Deckeneffekte vorfinden. Da das Auftreten von Boden- und Deckeneffekten ausgeschlossen werden kann, wird auf die Betrachtung der Itemschwierigkeit verzichtet. Denn abgesehen davon enthält der Silbentest mit 134 Items so viele Items, dass ein einzelner Schüler oder eine einzelne Schülerin während einer Minute wahrscheinlich ohnehin nicht alle Items lösen könnte, sodass die Itemschwierigkeit eines Items, welches weniger häufig getestet wurde als andere, eventuell unbegründet schwerer ausfällt, da nicht beantwortete Items im Sinne des Speedtests als falsch gelöste Items gewertet werden.

In der Jahrgangsstufe 4 streuen die einzelnen Ergebnisse ähnlich wie in der dritten Jahrgangsstufe sehr breit. Der geringste Wert aller Messzeitpunkte liegt bei 17 richtig gelesenen

Silben innerhalb einer Minute (MZP 1), der höchste Wert liegt bei 84 gelesenen Items (MZP 3) (s. Abbildung 15). Die größte Spannweite liegt mit einem Unterschied von 64 Items zur ersten Testung vor (s. Anhang, S. XXII).

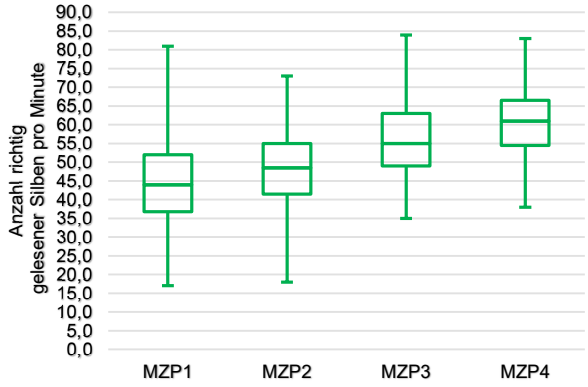


Abbildung 15: *Boxplot - Silben lesen (Jgst. 4)*

Bei Betrachtung der Boxplots (s. Abbildung 15) kann bereits ein Wachstum der Summenscores festgestellt werden. Median- und Minimalwerte steigen stark an, sodass der Median von der ersten bis zur letzten Testung um eine Anzahl von 17 korrekt gelesenen Items auf 61 richtig gelesene Silben innerhalb einer Minute angestiegen ist. Die Minimalwerte der einzelnen Messzeitpunkte steigen von anfänglich 17 richtig gelesenen Items um 21 Items auf 38 richtig gelesene Items in der vierten Testung an. Da außerdem der untere Quartilsabstand abnimmt und zum vierten Messzeitpunkt weit höher gelagert ist als zum ersten, kann insgesamt von einer Zunahme der Leistung in der Jahrgangsstufe 4 ausgegangen werden

Auch Tabelle 6 zeigt durch den steigenden Mittelwert in Zeile 3 den Leistungszuwachs in der vierten Jahrgangsstufe. Der Mittelwert der Summenscores aller Messzeitpunkte liegt bei M≈ 52,91. Dies entspricht in etwa 39,5% der Gesamtanzahl an Items in den Silbentests. Die Standardabweichung nimmt seit dem ersten Messzeitpunkt stetig ab, was auch bereits Abbildung 15 grob entnommen werden konnte.

Tabelle 6

Überblick über die Ergebnisse des Silbentests (Jgst. 4)

	MZP1	MZP2	MZP3	MZP4	M
N	24	38	40	35	34
M	44,25	48,71	55,85	60,03	52,91
Md	44	48,5	55	61	53
Mo	52	55	55	61	55
SD	12,94	11,43	10,4	10,12	12,52

Auch in der Testwertverteilung der Jahrgangsstufe 4 ist auffällig, dass zu Messzeitpunkt 3 der Mittel-, Median- und Modalwert zusammenfallen. Im Unterschied zur Jahrgangsstufe 3, ist es in dieser Stufe im dritten und nicht bei den ersten beiden Messzeitpunkten der Fall. Außerdem liegen die drei Werte in dieser Testung noch dichter beieinander (M≈ 55,85; Mo= 55; Md= 55). Demzufolge kann zu dem Messzeitpunkt 3, den Lageregeln nach zu urteilen, von einer symmetrischen Verteilung ausgegangen werden (vgl. Moosbrugger & Kelava 2012, S. 93). Auch zu Messzeitpunkt 4 liegen die drei Lageparameter (M≈ 60,03; Md= 61; Mo= 61) sehr nah beieinander.

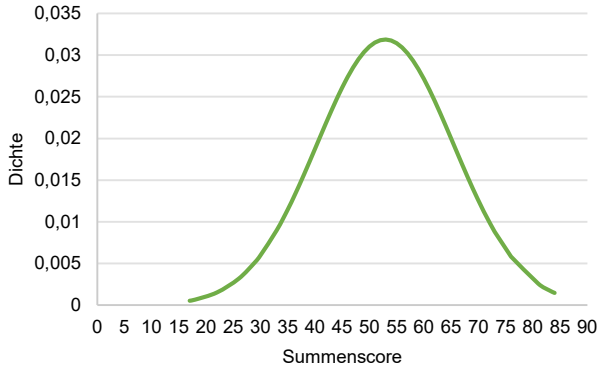


Abbildung 16: Normalverteilung – Silben (Jgst. 4)

Abbildung 16 stellt die Verteilung aller Messzeitpunkte dar. Wie in der Jahrgangsstufe 3 gilt auch hier, dass der Silbentest 134 Items und damit in keinem Fall die obere Grenze der maximal erreichbaren Summenscores erreicht wurde. Bei Berechnung des Schiefekoeffizienten erhält man mit (x) ≈ - 0,17 dennoch einen negativen Wert, welcher für eine linksschiefe Verteilung spricht. Nichtsdestotrotz kann das Auftreten von einem Boden- oder Deckeneffekt innerhalb des Leseflüssigkeitstests zum Silben lesen anhand der vorliegenden Werte ausgeschlossen werden.

Denn Maximalwerte von 84 richtig gelösten Items (≈ 62,7% der Gesamtanzahl an Items) in einem Test mit insgesamt 134 Items weisen genauso wenig auf Deckeneffekte hin, wie Minimalwerte von 17 Items (≈ 12,7%) auf Bodeneffekte. Abbildung 17 bestätigt diese Vermutung, da kein Testergebnis unter die Bodeneffektgrenze von 5% und ebenfalls keins über die Deckeneffektgrenze von 95% fällt. Auch wenn die Bodeneffektgrenze auf 10% angehoben werden würde, wird kein Testergebnis von Bodeneffekten betroffen sein. Lediglich bei einer Bodeneffektgrenze von 15% oder 20% würden ein bzw. bei 20% drei weitere Testergebnisse unter diese Grenze fallen. Der Maximalwert von 84 gelösten Items ist mit etwa 62,7% eindeutig zu weit unter der Deckeneffektgrenze, um Überlegungen zu starten, ob einige Probanden knapp unter dieser Grenze liegen könnten.

Bezüglich der einzelnen Itemschwierigkeiten wird wie aus den vorherigen Gründen auf eine Analyse verzichtet.

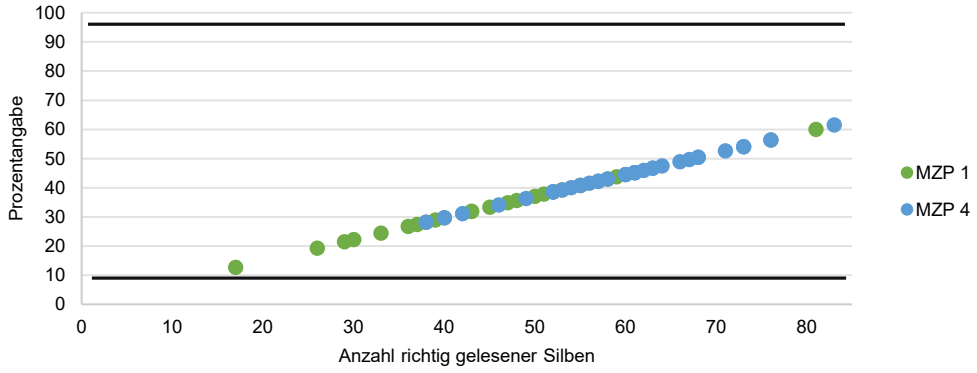


Abbildung 17: *Visualisierung der Ausprägung der Boden- und Deckeneffekte – Silben lesen (Jgst. 4)*

Testergebnisse des Leseflüssigkeitstests zum Wortlesen – N4

Der Leseflüssigkeitstest zum Wörterlesen umfasst genau wie der Test zum Leseverständnis eine Anzahl von 61 verschiedenen Items, von denen die SuS innerhalb einer Minute möglichst viele korrekt laut vorlesen müssen. Die Testergebnisse der Jahrgangsstufe 3 schwanken über alle Messzeitpunkte hinweg zwischen einem Minimum von 7 richtig gelesenen Items pro Minute und dem absoluten Maximum von 61 richtig gelesenen Items in dem genannten Zeitraum (s. Abbildung 18). Die größte Spannweite erreichen die einzelnen Summenscores zu Messzeitpunkt 1, da die Spannweite dort 53 beträgt und somit relativ nah an die Gesamtanzahl von 61 Items rückt. Durchschnittlich wurden in den Tests der Jahrgangsstufe 3 M= 33,79 Items richtig gelesen (s. Tabelle 7). Dies entspricht etwa 55,39% der Gesamtmenge an Items im Wortlesetest.

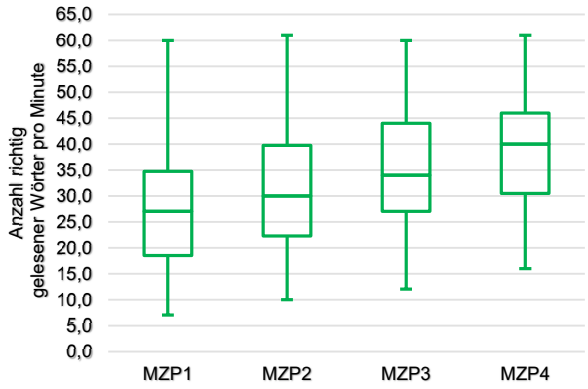


Abbildung 18: *Boxplot - Wörter lesen (Jgst. 3)*

Insgesamt betrachtet steigen Median, Mittelwert und Modalwert über alle vier Messzeitpunkte hin an, während die Standardabweichung relativ konstant bleibt (s. Tabelle 7). Durch die Betrachtung der Boxplots (s. Abbildung 18) kann bereits erschlossen werden, dass einige Kinder „an die Decke“ des Testes stoßen, da die Spannweite sowie die oberen Quartile abnehmen und ein Zuwachs wahrnehmbar ist. Insgesamt haben in den Messungen 2 und 4 bereits vier SuS der dritten Jahrgangsstufe die maximale Anzahl von 61 Items korrekt gelesen. Da die vier Summenscores von 61 nicht auf dieselben Probanden zurückzuführen sind, sondern von vier verschiedenen SuS erreicht wurde, wird die Tagesschwankungen der Schülerleistungen auch hier wieder bestätigt.

Tabelle 7

Überblick über die Ergebnisse des Wörterlesetests (Jgst. 3)

	MZP1	MZP2	MZP3	MZP4	
N	46	46	43	47	45,5
M	28,02	32,37	36	38,79	33,79
Md	27	30	34	40	32,5
Mo	32	21	34	41	32
SD	11,4	12,91	12,77	11,69	12,86

Die drei Maße der zentralen Tendenz fallen in diesen Testergebnissen zu keinem Messzeitpunkt zusammen, sodass man von einer nicht-symmetrischen Verteilung ausgehen kann. Lediglich bei der Betrachtung der gesamten Ergebnisse liegen die drei Parameter nah beieinander. Die Häufigkeitsverteilung in Abbildung 19 zeigt, dass nicht von einem rechtsgipfligen Verlauf, der einen Deckeneffekt bestätigen könnte, gesprochen werden kann, da die Verteilung mit einem Schwierigkeitskoeffizienten von (x) ≈ 0,32 rechtsschief ist. So liegt auch lediglich zu Messzeitpunkt 4 der Medianwert höher als der Mittelwert. Insgesamt betrachtet liegt, der rechtsschiefen Verteilung entsprechend, der Median mit Md= 32,5 jedoch unter dem Mittelwert mit M≈ 33,79. Da die Verteilung weder rechts noch links zensiert wird, kann eine starke Ausprägung von Boden- und Deckeneffekten ausgeschlossen werden.

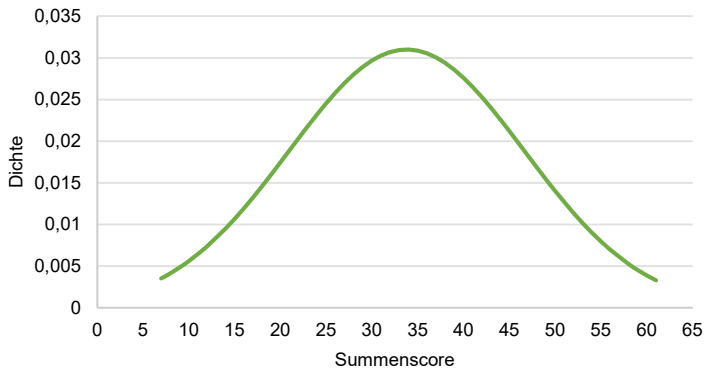


Abbildung 19: *Normalverteilung - Wörter lesen (Jgst. 3)*

Die Grenzen der Decken- und Bodeneffekte werden wie im Test zum Silben lesen bei 95% (≈ 58 richtige Items) und 5% (≈ 3 richtige Items) gesetzt (s. Abbildung 20). Zum ersten Messzeitpunkt lag lediglich eine Schülerin (≈ 2,17%) mit 60 richtig gelösten Items über der Grenze des Deckeneffekts. Bereits zu diesem Zeitpunkt lag kein Ergebnis unter der Bodeneffektgrenze. Denn die geringsten Testwerte lagen bei 7 (≈ 11,48%) und 8 (≈ 13,11%) richtig gelösten Items. Damit würden erst dann Bodeneffekte entstehen, wenn diese Grenze bei 15% angesetzt werden würde. Hinsichtlich der Bodeneffekte verändert sich auch während der nächsten drei Testungen nichts. Jedoch steigen die unteren Testergebnisse an, sodass der geringste Wert während des dritten Messzeitpunktes bei 12 richtig gelösten Items (≈ 19,67%) liegt und während des vierten Messzeitpunktes mit einem Summenscore von 16 (≈ 26,23%) sogar eine 25%-Bodeneffektgrenze überschreiten würde.

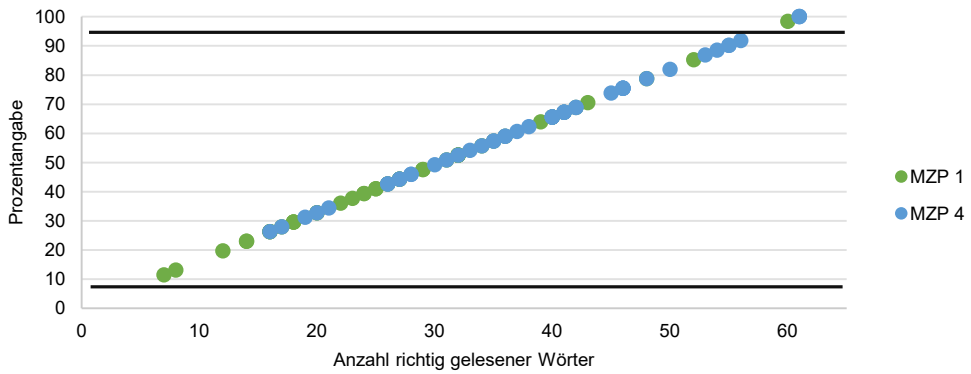


Abbildung 20: *Visualisierung der Ausprägung der Boden- und Deckeneffekte – Wörter lesen (Jgst. 3)*

Da bereits festgestellt wurde, dass die niedrigsten Summenscores stets angestiegen sind, lässt sich vermuten, dass auch die oberen Testwerte stetig zugenommen haben. Lag zu Testzeitpunkt 1 lediglich ein Proband (≈ 2,17%) über der 95%-Grenze, waren es in der zweiten Messung bereits zwei SuS (≈ 4,34%) und in der dritten Testdurchführung fünf SuS

(≈ 11,63%), deren Summenscores über 95% betragen. Während des vierten Testzeitpunktes sank die Anzahl der Summenscores, die Deckeneffekte aufweisen, sodass zu diesem Zeitpunkt lediglich die Summenscores von drei SuS (≈ 6,38%) Deckeneffekte aufgewiesen haben. Diese drei Probanden erzielten jedoch allesamt 100%. Daraus ergibt sich ein Gesamtbild von 11 Probanden (≈ 6,04%), die in der Jahrgangsstufe 3 im Wörterlesetest einen Deckeneffekt aufweisen. Wird die Grenze breiter gesteckt und geschaut wie viele SuS 90% (≈ 55 richtige Items) der Items korrekt vorlesen, um zu sehen, ob einige Probanden knapp unter der Deckeneffektgrenze liegen, erhält man bei Berücksichtigung aller Messzeitpunkte eine Anzahl von 16 SuS (≈ 8,79%) über der 90%-Grenze.

Von den sieben SuS mit Migrationshintergrund schneiden vier SuS bei einem Mittelwert von M≈ 28,02 durchschnittlich bis überdurchschnittlich ab. Die anderen drei SuS, darunter der Schüler, dem der Förderschwerpunkt Deutsch zugeschrieben wurde, erreichen Summenscores, die in etwa eine Standardabweichung schlechter ausfallen. Alle Summenscores dieser sieben SuS verbessern sich über die vier Messzeitpunkte stark, sodass eine Schülerin in der vierten Testung den maximalen Summenscore von 61 erreicht. Die Summenscores der anderen sechs SuS mit Migrationshintergrund liegen zum vierten Messzeitpunkt zwischen 32 und 42 richtig gelösten Items und liegen damit nahe an dem Mittelwert M≈ 38,79.

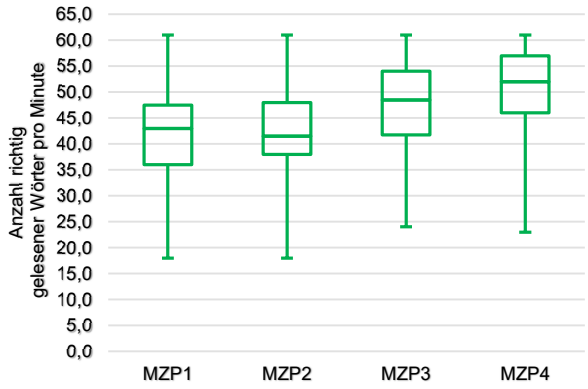


Abbildung 21: *Boxplot - Wörter lesen (Jgst. 4)*

Die Testergebnisse der Jahrgangsstufe 4 im Wörterlesen sind ähnlich zu denen der Jahrgangsstufe 3. Die Minimalwerte der Testung liegen im Messzeitpunkt 1 und 2 bei 18 richtig gelösten Items (s. Abbildung 21). Das bedeutet sogar die geringsten Summenscores aller vier Messzeitpunkte erreichen 29,51% der Gesamtmenge an Items. Während des dritten und vierten Testtermins lagen die minimalen Werte bei 24 (≈ 39,34%, MZP 3) und bei 23 (≈ 37,7%, MZP 4), also ebenfalls sehr hoch. Maximal erreichen die SuS auch hier die Höchstpunktzahl, den Summenscore von 61 Items und zwar während allen vier Testungen. Da der Minimalwert in der Jahrgangsstufe 4 jedoch bei 18 und nicht wie in der Jahrgangsstufe 3 bei

einem Summenscore von 7 liegt, erreichen die Testergebnisse während des ersten und des zweiten Messzeitpunktes lediglich eine Spannweite von 43. Der abnehmende obere Quartilsabstand sowie der steigende Median im Messzeitpunkt drei und vier weisen bereits auf Deckeneffekte hin. Denn so erreichen beispielsweise zu Messzeitpunkt 4, ablesbar anhand des Medians (Md= 52), 50% der Probanden einen Summenscore von 52 oder mehr und 25% der SuS sogar einen Summenscore von 57 oder mehr. Zur Übersicht liegen im Anhang ab S. XXIV die Werte aller Messzeitpunkte vor.

Tabelle 8

Überblick über die Ergebnisse des Wörterlesetests (Jgst. 4)

	MZP1	MZP2	MZP3	MZP4	
N	39	38	40	37	38,5
M	41,41	41,81	49,93	49,65	44,92
Md	43	41,5	48,5	52	45,5
Mo	44	42	43	56	42
SD	10,84	10,65	9,16	9,93	10,73

Die Standardabweichung (s. Tabelle 8) fällt besonders im Hinblick auf die gesamte Standardabweichung aller vier Messzeitpunkte in der Jahrgangsstufe 4 mit SD= 10,73 geringer aus als in der Jahrgangsstufe 3 (SD= 12,86). In der Jahrgangsstufe 4 ist außerdem ein Sinken der Standardabweichung zu verzeichnen; mit Ausnahme des vierten Testzeitpunktes, indem sie wieder leicht zunahm.

Auffällig ist, dass der Mittelwert der Summenscores von dem ersten zum zweiten Messzeitpunkt sowie von der dritten zur vierten Testung kaum bzw. gar nicht ansteigt und somit lediglich vom zweiten zum dritten Messzeitpunkt eine positive Veränderung der durchschnittlichen Summenscores wahrgenommen werden kann (s. Tabelle 8). Die Median- (s. Abbildung 21 oder Anhang, S. XXV) und Modalwerte (s. Tabelle 8) bleiben während der ersten drei Messungen relativ konstant und steigen erst zum vierten Messzeitpunkt rasant an. Dies bezüglich ist auch auffällig, dass sich zum zweiten Messzeitpunkt die Werte des Medians (Md= 41,5), des Mittelwertes (M= 41,81) und des Modalwertes (Mo= 42) nahezu decken, was bedeutet, dass eine symmetrische Verteilung vorliegt (vgl. Moosbrugger & Kelava 2012, S. 93).

Ähnlich zu den Ergebnissen des sinnentnehmenden Lesetests lassen die Betrachtung der Boxplots (s. Abbildung 21), der hohen Mittelwerte der Summenscores sowie die abnehmende Standardabweichung (s. Tabelle 8) Deckeneffekte vermuten. Auch die Betrachtung der Kurve der Normalverteilung der Messergebnisse spricht für das Auftreten eines Deckeneffektes, da die Kurve rechts zensiert ist (s. Abbildung 22). Außerdem fällt bereits anhand der aufgeführten Abbildungen auf, dass die Verteilung rechtsgipflig ist, da ein Großteil der Probanden mehr als die Hälfte der Items löst, in diesem konkreten Fall sogar mehr als zwei Drittel, der im Wortlesetest angebotenen Items korrekt (vgl. Lienert & Raatz 1998, S. 155). Auch der Schiefekoeffizient von (x) ≈ -0,47 beweist eine linksschiefe Verteilung. Dem entspricht auch, dass der Medianwert während allen Messzeitpunkten, abgesehen von Messzeitpunkt 2, über dem Mittelwert liegt (MZP 1: Md= 43, M≈ 41,41; MZP 2: Md= 38, M≈ 41,82; MZP 3: Md= 48,5, M≈ 46,93; MZP 4: Md= 52, M≈ 49,65).

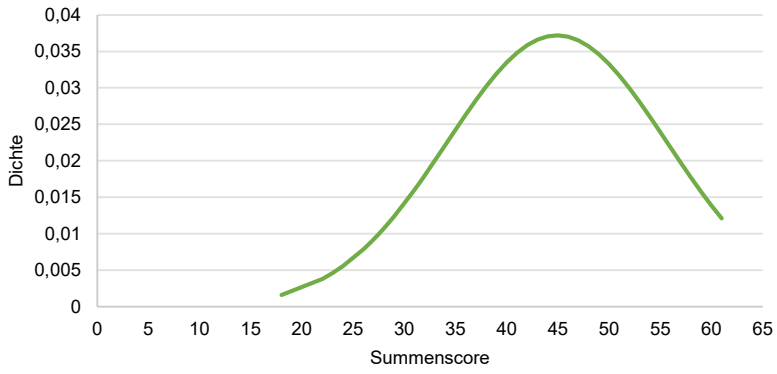


Abbildung 22: *Normalverteilung – Wörter lesen (Jgst. 4)*

Abbildung 23 zeigt, dass das Auftreten von Bodeneffekten innerhalb dieser Testungen definitiv ausgeschlossen werden kann, da bis auf einen Schüler, der während des ersten und auch des zweiten Messzeitpunktes einen Summenscore von 18 richtig gelösten Items (≈ 29,51% der Gesamtanzahl an Items) erreichte, somit aber definitiv auch nicht unter der Bodeneffektgrenze liegt, alle SuS mehr als 35% der Items korrekt vorlesen konnten. Im Hinblick auf mögliche Deckeneffekte wird bereits zum ersten Messzeitpunkt offensichtlich, dass eine Schülerin die Höchstpunktzahl erreichte. Insgesamt lagen mit diesem Probanden n= 3 SuS (≈ 7,69%) im ersten Messzeitpunkt über der gesetzten Grenze von 95%. Während der folgenden Testungen stieg diese Anzahl stetig an, so waren es n= 4 SuS (≈ 10,53%) zu Testzeitpunkt 2, n= 5 Probanden (≈ 12,5%) in der dritten Testung und n= 9 SuS (≈ 24,32%) zu Messzeitpunkt 4. Insgesamt lagen während allen Messzeitpunkten also 21 (≈ 13,64%) SuS über der Deckeneffektgrenze von 95%, davon erreichten 7 Probanden (≈ 4,55%) den höchstmöglichen Summenscore von 61 Items.

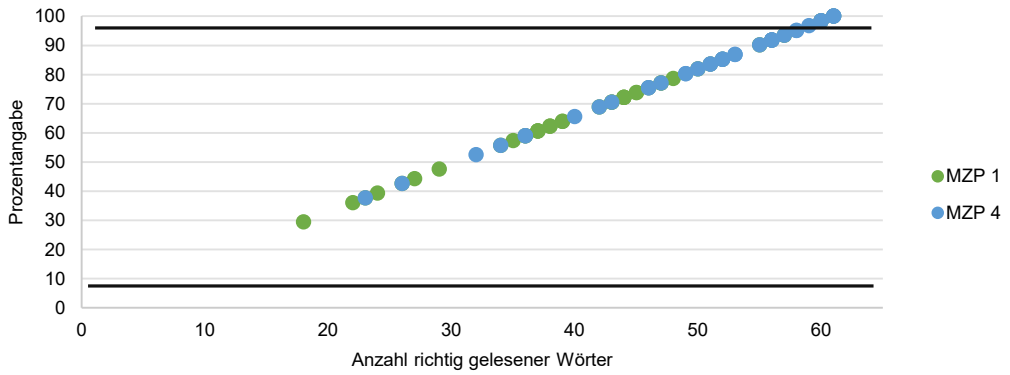


Abbildung 23: *Visualisierung der Ausprägung der Boden- und Deckeneffekte – Wörter lesen (Jgst. 4)*

Wie auch bereits bei den zuvor vorgestellten Testergebnissen ist auch hier auffällig, dass viele der Probanden knapp unter den Grenzwert von 58 richtig gelösten Items (≈ 95%) fallen. Bezieht man daher auch die SuS mit ein, die 55 Items (≈ 90%) während einer Minute richtig vorgelesen haben, so liegen über alle vier Messzeitpunkte hinweg 34 von insgesamt 154 Testergebnissen (≈ 22,08%) über der herabgesetzten Deckeneffektsgrenze von 90% (MZP 1: n= 6 (≈ 15,38); MZP 2: n= 5 (≈ 13,16%), MZP 3: n= 8 (≈ 20,0%), MZP 4: n= 15(≈ 20,54%)). Die gesetzte Grenze des Bodeneffekts zu verschieben, ergibt in diesem Fall keinen Sinn, da wie bereits erwähnt, bis auf eine Ausnahme, alle SuS mehr als 35% der Items richtig lösen konnten und eine bei 35% angesetzte Bodeneffektgrenze definitiv zu hoch liegen würde.

Im Hinblick auf die Ergebnisse der vier SuS mit Migrationshintergrund kann festgestellt werden, dass sie während der ersten Testung noch eindeutig unter dem Mittelwert liegen, sich diesem aber in den weiteren Messungen immer weiter annäherten oder auch überholten. Die Ausnahme stellt in diesem Fall die Schülerin dar, der neben dem Migrationshintergrund auch der Förderschwerpunkt Deutsch zugeschrieben wurde, d.h. aktuell noch die deutsche Sprache erlernt. Denn sie erreicht während allen Messzeitpunkten unterdurchschnittliche Leistungen, wie auch ihr durchschnittlicher Summenscore in Höhe von M= 33 im Vergleich zu dem Mittelwert aller Summenscores in Höhe von M≈ 44,92 beweist. Bezüglich der anderen beiden SuS, denen ein Förderschwerpunkt zugeschrieben wurde, fällt auf, dass der Schüler mit diagnostiziertem Förderschwerpunkt Körperliche und motorische Entwicklung während allen Messzeitpunkten über dem Durchschnitt der gesamten Jahrgangsstufe liegt. So liegt auch sein durchschnittlicher Summenscore von M= 49,75 über dem Mittelwert der gesamten Stufe (M≈ 44,92). Der Schüler mit zugeschriebenem Förderschwerpunkt Lernen zeigt durchgängig leicht unterdurchschnittliche Leistungen, so liegt auch sein durchschnittlicher Summenscore in Höhe von M= 39,75 unter dem Durchschnitt der gesamten Jahrgangsstufe (M≈ 44,92).

Insgesamt kann also davon ausgegangen werden, dass auch der Leseflüssigkeitstest zum Wörter lesen auf der Niveaustufe 4 in der dritten und insbesondere in der vierten Jahrgangsstufe Deckeneffekte aufweist.

Um auch für dieses Testformat eine mögliche Begründung für das Auftreten der Deckeneffekte zu finden, wird nun die Schwierigkeit der einzelnen Items betrachtet. Da dieses Testformat, genau wie der sinnentnehmende Lesetest über 61 Items verfügt, bietet sich die Schwierigkeitsanalyse der Items hier mehr an, da die SuS meist mehr als die Hälfte der Items lösen. Der durchschnittliche Schwierigkeitsindex aller Items zu allen Messzeitpunkten im Wörterlesetest in der Jahrgangsstufe 4 liegt bei $P_i \approx 73{,}69$. Ein Schwierigkeitsindex von $P_i \approx 73{,}69$ liegt im optimalen Bereich, kann jedoch als eher leicht eingeschätzt werden. Über alle Messzeitpunkte hinweg haben sich einige Items als konstant leichte Items erwiesen. Dazu gehören beispielsweise die Items „Wolke“ ($P_i \approx 91{,}61$), „Graben“ ($P_i \approx 87{,}61$), „Gleis“ ($P_i \approx 87{,}71$), und „Traube“ ($P_i \approx 86{,}29$). Dagegen stellen sich die Items „Puls“ ($P_i \approx 57{,}07$), „falsche“ ($P_i \approx 51{,}96$) und „Furcht“ ($P_i \approx 32{,}67$) als die in dieser Longitudinalstudie am seltensten korrekt gelösten Items heraus (s. Anhang, S. XXXIII). Itemschwierigkeiten um $P_i \approx 50$ sind aufgrund ihrer mittelschweren Komplexität jedoch wünschenswert (vgl. Schmidt-Atzert & Amelang 2012, S. 114). Bei Begutachtung der Tabelle 9 gilt es wieder zu berücksichtigen, dass die Schwierigkeitsindizes des ersten Messzeitpunktes aufgrund der immer gleichen Reihenfolge der Tests anders zu bewerten sind.

In der dritten Jahrgangsstufe liegt ein durchschnittlicher Schwierigkeitsindex von $P_i \approx 55{,}07$ vor. Der Unterschied des Schwierigkeitsindizes zur vierten Jahrgangstufe lässt sich dadurch erklären, dass die SuS der Jahrgangsstufe 3 insgesamt weniger Items korrekt gelesen haben, sodass die Itemschwierigkeit schlechter ausfallen muss. In der Jahrgangsstufe 3 konnten keine Items ausfindig gemacht werden, die über alle Messzeitpunkte konstant leicht abgeschnitten haben. Jedoch ist auffällig, dass dieselben Items wie in der vierten Jahrgangsstufe konstant schwerer ausfallen (s. Tabelle 9). So sind die Items *„falsche“* mit einem durchschnittlichen Schwierigkeitsindex von $P_i \approx 25{,}1$ in der Jahrgangsstufe 3 und von $P_i \approx 51{,}96$ in der Jahrgangsstufe 4 und das Item „Furcht“ (Jgst. 3: $P_i \approx 18{,}67$; Jgst. 4: $P_i \approx 32{,}67$) in beiden Jahrgangsstufen, die am seltensten gelösten Items. Da auch hier berücksichtigt werden muss, dass die Itemreihenfolge in der Messung festgelegt ist, ist es von hoher Relevanz zu erwähnen, dass das Item „Furcht“ in der ersten Testung das drittletzte der 61 Items darstellt, während sich das Item „falsche“ in der hinteren Mitte der Testreihenfolge befindet, sodass der extrem hohe Schwierigkeitsgrad bzw. der extrem niedrige Schwierigkeitsindex des Items „Furcht“ zu Testzeitpunkt 1 revidiert werden muss.

Tabelle 9

Exemplarische Itemschwierigkeiten des Wörterlesetests

Jahrgangsstufe 3				
Items	MZP 1	MZP 2	MZP 3	MZP 4
	P_i	P_i	P_i	P_i
falsche	15,22	28,26	25	31,91
Furcht	2,17	26,09	27,27	19,15

Jahrgangsstufe 4				
Items	MZP 1	MZP 2	MZP 3	MZP 4
	P_i	P_i	P_i	P_i
Wolke	100	86,84	85	94,59
Graben	94,87	73,68	90	91,89
Gleis	100	78,95	80	91,89
Traube	100	81,58	82,5	81,08
Puls	66,67	47,37	57,5	56,76
falsche	51,28	42,11	55	59,46
Furcht	7,69	36,84	37,5	48,65

Testergebnisse des Leseflüssigkeitstests zum Pseudowörterlesen – N4

Der Test zum Pseudowörterlesen ist mit einer Anzahl von 189 Items der „item-umfangreichste“ Test der Levumi-Plattform. Der geringste Summenscore aller vier Messungen liegt bei 5 richtig gelesenen Pseudowörtern innerhalb einer Minute, der höchste Wert wurde in Messzeitpunkt 4 mit einer Anzahl von 45 richtigen Items pro Minute erreicht. Mit einer Spannweite von 38, besteht die größte Spannweite zu Messzeitpunkt 4. Allein aufgrund von Abbildung 24 kann ein immenser Unterschied zu den anderen Tests festgestellt werden. Denn

den Boxplots der drei anderen durchgeführten Tests konnte immer bereits bei bloßer Betrachtung ein starker Anstieg von Minimal- und Maximalwerten, der Mediane oder der gesamten Boxen abgelesen werden. Im Pseudowörtertest steigt der Maximalwert von einem anfänglichen Summenscore von 36 zu Messzeitpunkt 1 auf 45 zu Messzeitpunkt 4 an, der Minimalwert der Testungen steigt jedoch nur um 2 gelöste Items an, d.h. von einem Minimalwert von 5 in Testung 1 zu einem Minimalwert von 7 in Testung 4. Im Gegensatz zu den anderen Testergebnissen, in denen die Box, d.h. der Abstand zwischen dem oberen und dem unteren Quartil entweder relativ konstant blieb oder sich verringerte, variiert dies in der Pseudoworttestung der Jahrgangsstufe 3 stärker als in den anderen Testformaten.

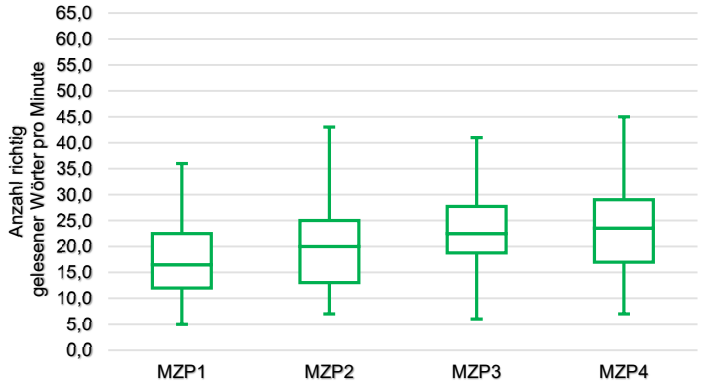


Abbildung 24: *Boxplot - Pseudowörter lesen (Jgst. 3)*

Außerdem bleiben die Medianwerte der einzelnen Testungen, wenn die hohe Itemzahl bedacht wird, relativ konstant. Liegt der Median zu Messzeitpunkt 1 bei Md= 16,5, so ist dieser zum vierten Messzeitpunkt (Md= 23,5) nicht stark angestiegen (s. Abbildung 24 oder Anhang, S. XXVII).

Gleiches gilt für den Mittelwert (s. Tabelle 10). Steigt dieser von Messzeitpunkt 1 zu Messzeitpunkt 2 und von diesem zum dritten Messzeitpunkt um durchschnittlich jeweils drei mehr gelesene Items an, nimmt er zur vierten Testung hin wieder leicht ab. Von Messzeitpunkt 1 zu Messzeitpunkt 4 wurde die durchschnittlich gelöste Itemanzahl in der Jahrgangsstufe 3 folglich nur um sechs Items verbessert. Im Vergleich zu den Ergebnissen der Jahrgangsstufe 3 im Silben- und im Wörterlesen stellt dies einen recht geringen Zuwachs dar. Denn innerhalb des Tests zum Silbenlesen nahm die durchschnittliche Anzahl korrekt gelesener Items um 13 Items zu und innerhalb des Tests zum Wörterlesen immerhin um 10 Items. Auffällig ist auch der über alle Messzeitpunkte hinweg schwankende Modalwert sowie die stetig ansteigende Standardabweichung.

Tabelle 10

Überblick über die Ergebnisse des Pseudowörterlesetests (Jgst. 3)

	MZP1	MZP2	MZP3	MZP4	M
N	46	45	44	44	44,75
M	17,17	20,11	23,32	23,16	20,89
Md	16,5	20	22,5	23,5	20
Mo	20	13	19	17	19
SD	6,97	8,32	8,37	8,75	8,5

Da Median-, Modal- und Mittelwert zu keinem Testzeitpunkt gleich ausfallen, herrscht eine nicht-symmetrische Verteilung. Da der Modalwert auch lediglich zu Messzeitpunkt 1 über dem Mittelwert liegt und der Mittelwert in den übrigen Messzeitpunkten weit über dem Modalwert liegt, kann auch eine rechtsgipflige Verteilung ausgeschlossen werden. Dies fällt auch bei der Betrachtung von Abbildung 25 auf, indem die Verteilung eher linksgipflig wirkt. Wird dies mit der Berechnung des Schiefekoeffizienten überprüft, wird mit einem Wert von (x) ≈ 0,43 eine rechtsschiefe Verteilung bestätigt.

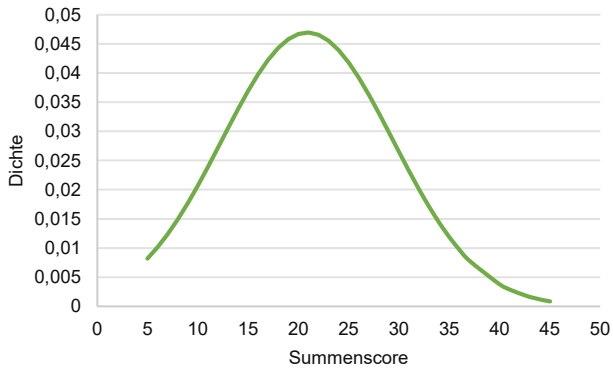


Abbildung 25: *Normalverteilung – Pseudowörter lesen (Jgst. 3)*

Wie Abbildung 26 zu entnehmen ist, werden auch in diesem Test die Grenzen der Boden- und Deckeneffekte bei 5% bzw. bei 95% angesetzt. Da mit einer maximalen Anzahl von 45 korrekt gelösten Items (≈ 23,81%), die gesamte Itemanzahl von 189, bzw. die Deckeneffektgrenze von ca. 180 Items (≈ 95%) bei Weitem nicht erreicht wird, kann das Auftreten von Deckeneffekten im Pseudoworttest in der dritten Jahrgangsstufe definitiv ausgeschlossen werden. Der für diese Testung maximale Wert von 45 Items wurde lediglich von zwei SuS erreicht und ist, wie ebenfalls Abbildung 26 zu entnehmen ist, mit gewissem Abstand der einzige derartig hohe Wert.

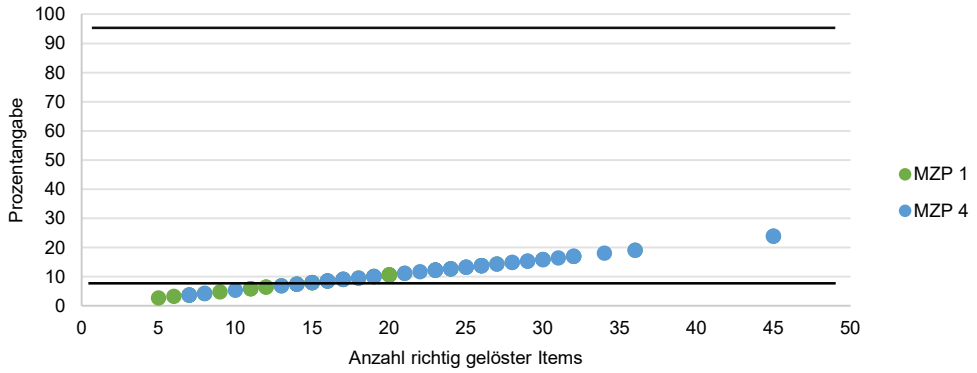


Abbildung 26: *Visualisierung der Ausprägung der Boden- und Deckeneffekte – Pseudowörter lesen (Jgst. 3)*

Da auch hier die Grenze des Bodeneffekts bei 5% angelegt wurde, müssen die SuS mindestens 9 Items innerhalb einer Minute richtig vorlesen. Damit liegen während allen vier Messzeitpunkten Bodeneffekte vor, insbesondere gilt dies für den ersten Messzeitpunkt, wie Abbildung 26 verdeutlicht. So liegen während der ersten Testung 8 Probanden (≈ 18,18%) unter der genannten Grenze. Die Anzahl der Teilnehmer, die unter diese Grenze fallen, nimmt mit den Messzeitpunkten ab, sodass zum zweiten Messzeitpunkt nur noch die Testergebnisse von vier SuS (≈ 8,89%), zum dritten nur noch die Ergebnisse zweier SuS (≈ 4,55%) und zum vierten Messzeitpunkt dann aber wieder die Testergebnisse von drei SuS (≈ 6,82) unter dieser Grenze liegen. Insgesamt werden nach den vorliegenden Daten die Testergebnisse von 17 SuS (≈ 9,5% aller Testergebnisse) durch das Auftreten des Bodeneffekts eingeschränkt. Die Relevanz der hier ermittelten Bodeneffekte wird im anschließen Kapitel weiter diskutiert.

Wird in den Blick genommen, ob während allen Messzeitpunkten dieselben SuS durch den Bodeneffekt eingeschränkt werden, oder ob die Leistungen der SuS so stark schwanken, dass immer andere SuS unter diese Grenze fallen, wird offensichtlich, dass alle SuS, die zu den späteren Messungen unter der genannten Grenze liegen, auch bereits zum ersten Messzeitpunkt weniger als 5% der Items richtig lösen konnten. Davon liegen die Ergebnisse eines Schülers konstant zu allen vier Messterminen unter der Grenze, bezüglich seines individuellen Lernverlaufs ist auffällig, dass er keine Lernfortschritte macht, sondern seine Leistung rückläufig und stagnierend ist (MZP 1: Summenscore 9 (≈4,76%); MZP 2: Summenscore 7 (≈ 3,7%); MZP 3: Summenscore 6 (≈ 3,17%); MZP 4: Summenscore 8 (≈4,23%)). Ursachen dafür können in der dem Untersuchungsteam zur Verfügung gestellten Schülerbiographie nicht gefunden werden. Die einzige Auffälligkeit ist sein mit 11 Jahren relativ hohes Alter für die Jahrgangsstufe 3, was möglicherweise auf Zurückstufungen oder ähnliches hindeuten kann.

In der vierten Jahrgangsstufe lässt sich ein ähnliches Bild wie in den Ergebnissen des Pseudoworttests der Jahrgangsstufe 3 vorfinden. Wie Abbildung 27 bereits zeigt variieren die Ergebnisse der einzelnen Messzeitpunkte sehr stark. Der Minimalwert aller Messzeitpunkte liegt hier zu Messzeitpunkt 1 mit einer Anzahl von 7 gelösten Items vor. Die Minimalwerte der weiteren Messungen steigen bis auf einen Summenscore von 11 zu Messzeitpunkt 4 an. Der Höchstwert liegt zu Messzeitpunkt 3 mit einer gelösten Itemanzahl von 50 vor, zu Messzeitpunkt 4 sank das Maximum lediglich um ein Item auf einen derzeit maximalen Summenscore von 49 ab.

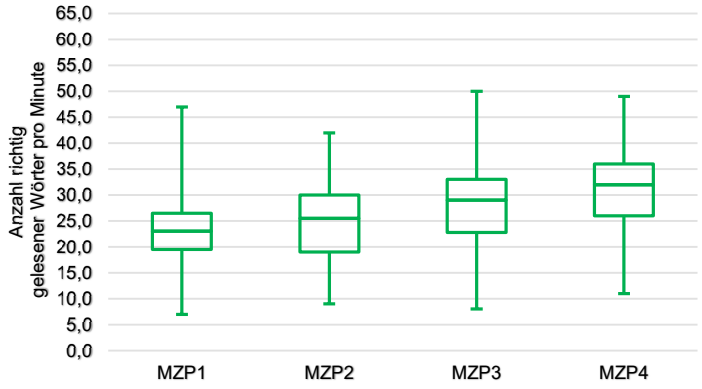


Abbildung 27: *Boxplot - Pseudowörter lesen (Jgst. 4)*

So liegt auch die größte Spannweite der Testergebnisse mit einem Unterschied von 42 Items während Messzeitpunkt 3 vor. Anders als in der Ergebnisdarstellung zum Pseudowörtertest in der Jahrgangstufe 3 kann hier bereits in Abbildung 27 abgelesen werden, dass neben den Medianwerten von Md=23 in der ersten Testung zu Md= 32 in der vierten Testung, auch die gesamten Boxen sichtlich ansteigen. Das bedeutet, dass die gesamten Schülerleistungen stark ansteigen.

Tabelle 11

Überblick über die Ergebnisse des Pseudowörterlesetests (Jgst. 4)

	MZP1	MZP2	MZP3	MZP4	M
N	39	38	40	37	38,5
M	23,33	24,78	28,48	31,3	26,94
Md	23	25,5	29	32	27
Mo	26	19	33	28	19
SD	6,59	7,67	7,9	8,47	8,28

Tabelle 11 zeigt, dass ähnlich wie die Medianwerte auch die Mittelwerte zu allen Messzeitpunkten ansteigen. Im Vergleich zu den Testergebnissen der Jahrgangsstufe 3 liegt der Mittelwert aller Messzeitpunkte in der Jahrgangsstufe 4 bei M≈ 26,94 und damit um einiges höher als der Mittelwert der Jahrgangsstufe 3 (M≈ 20,89). Ähnlich zu den Testergebnissen

der Jahrgangsstufe 3 schwankt der Modalwert auch in der Jahrgangsstufe 4 sehr stark und liegt zu Messzeitpunkt 4 nicht viel höher als zum ersten Messzeitpunkt.

Auch in dieser Jahrgangsstufe fallen Median-, Modal- und Mittelwert zu keinem Messzeitpunkt zusammen, sodass von einer nicht-symmetrischen Verteilung ausgegangen werden kann. Da lediglich zu Messzeitpunkt 1 und zu Messzeitpunkt 3 der Modalwert über dem Mittelwert liegt, in der gesamten Studie aber der Mittelwert M≈ 26,94 über dem Modalwert von Mo= 19 liegt, spricht dies für eine rechtsschiefe Verteilung aller Summenscores. So liegt auch der Schiefekoeffizient mit (x) ≈ 0,2 im positiven Bereich und bestätigt somit die vermutete rechtsschiefe Verteilung der Summenscores.

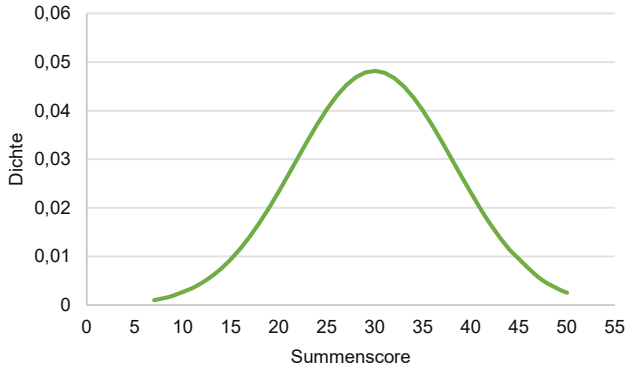


Abbildung 28*: Visualisierung der Ausprägung der Boden- und Deckeneffekte – Pseudowörter Lesen (Jgst. 4)*

Abbildung 29 zeigt nun, dass von Messzeitpunkt 1 zu Messzeitpunkt 4 ein Zuwachs stattgefunden haben muss. Auch im Vergleich zu Abbildung 26 der Jahrgangsstufe 3 kann ein Leistungsunterschied zwischen der dritten und vierten Jahrgangsstufe vermutet werden. Bezüglich der Boden- und Deckeneffekte werden hier die üblich angewandten Grenzen angewendet. Folglich liegt ein Bodeneffekt vor, wenn 5% oder weniger (≈ max. 9 Items) gelöst werden; ein Deckeneffekt liegt demnach vor, wenn 95% oder mehr (≈ mind. 180 Items) richtig vorgelesen werden. Mit einer maximal gelösten Anzahl von 50 korrekt vorgelesenen Items (≈ 26,46%) liegen die Summenscores der SuS auch hier weit unter der genannten Deckeneffektgrenze.

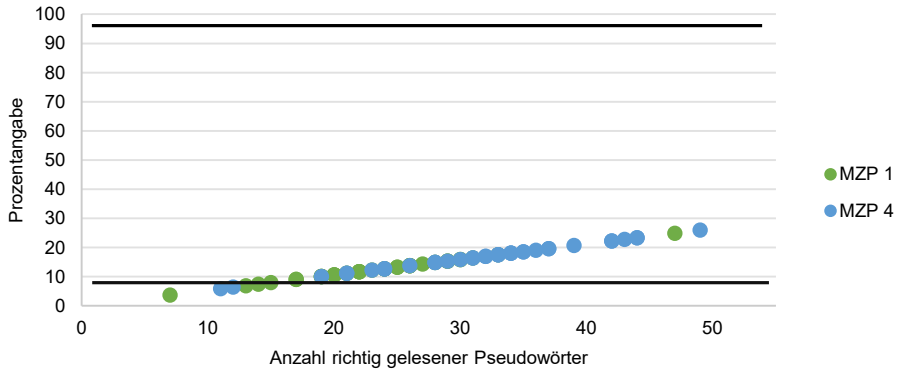


Abbildung 29: *Visualisierung der Ausprägung der Boden- und Deckeneffekte – Pseudowörter lesen (Jgst. 4)*

Hinsichtlich der Bodeneffektgrenze lässt sich feststellen, dass während der ersten Testung ein einziges Testergebnis (7 richtig gelöste Items ≈ 3,7%) unter der Grenze liegt, ebenso zu Testzeitpunkt 2. Dieses Ergebnis liegt mit neun richtig gelesenen Items direkt auf der Grenze. Während zu Messzeitpunkt 3 ebenfalls noch ein Testwert (Summenscore von 4 korrekt gelesenen Items ≈ 4, 23%) unter dieser Grenze lag, fällt zu Messzeitpunkt 4 kein Ergebnis mehr unter die Grenze des Bodeneffektes. Insgesamt werden in der Jahrgangsstufe 4 also die Testergebnisse von nur drei Probanden (≈ 1,95% aller Testergebnisse der Jahrgangsstufe 4 im Pseudowort lesen) durch den Bodeneffekt begrenzt. Wie schon in Bezug auf die Jahrgangsstufe 3 erwähnt, wird das Auftreten der Bodeneffekte im Pseudowörter lesen im anschließenden Kapitel weiter diskutiert.

Tabelle 12 gibt in der Ergebnisdarstellung einen abschließenden Überblick über die aufgetretenen Boden- und Deckeneffekte sowie ihrer Ausprägung in Abhängigkeit der Jahrgangsstufe und des Testformats innerhalb dieser Langzeitstudie. Die grau markierten Fächer verdeutlichen, dass zu diesen Messzeitpunkten oder auch im gesamten Silbentest kein Boden- oder Deckeneffekt aufgetreten ist.

Tabelle 12

Gesamtdarstellung der aufgetretenen Boden- und Deckeneffekte in der mit der Plattform Levumi durchgeführten Langzeitstudie

		Jahrgangsstufe 3					Jahrgangsstufe 4				
		MZP 1	MZP 2	MZP 3	MZP 4	insgesamt	MZP 1	MZP 2	MZP 3	MZP 4	insgesamt
Sinnentnehmendes Lesen	Bo	9 Pbn (≈ 20,45%)	5 Pbn (≈ 10,87%)	4 Pbn (≈ 9,09%)	1 Pbn (≈ 2,38%)	19 Pbn (≈ 10,8%)					
	De	1 Pbn (≈ 2,27%)	2 Pbn (≈ 4,35%)	5 Pbn (≈ 11,36%)	9 Pbn (≈ 21,43%)	17 Pbn (≈ 9,66%)	6 Pbn (≈ 15,79%)	5 Pbn (≈ 15,15%)	14 Pbn (≈ 35%)	22 Pbn (≈ 55%)	47 Pbn (≈ 31,13%)

Silben lesen	Bo										
	De										
Wörter lesen	Bo										
	De	1 Pbn (≈ 2,17%)	2 Pbn (≈ 4,34%)	5 Pbn (≈ 11,63%)	3 Pbn (≈ 6,38%)	11 Pbn (≈ 6,04%)	3 Pbn (≈ 7,69%)	4 Pbn (≈ 10,53%)	5 Pbn (≈ 12,5%)	9 Pbn (≈ 24,32%	21 Pbn (≈ 13,64%)
Pseudowörter lesen	Bo	8 Pbn (≈ 19,57%)	4 Pbn (≈ 8,88%)	2 Pbn (≈ 4,55%)	3 Pbn (≈ 6,82%)	17 Pbn (≈ 9,5%)	1 Pbn (≈ 2,56%)	1 Pbn (≈ 2,63%)	1 Pbn (≈ 2,5%)		3 Pbn ≈ 1,95%
	De										

Anmerkung: Abkürzung für Bodeneffekt in der Tabelle = Bo

Abkürzung für Deckeneffekt in der Tabelle = De

Abkürzung für Probanden = Pbn

7 Interpretation und Diskussion der Forschungsergebnisse

Die in dieser Langzeitstudie zugrunde gelegte Fragestellung, wie stark Boden- und Deckeneffekte in den Tests der Lernplattform Levumi innerhalb der Lerndomäne „Lesen“ in den Klassenstufen 3/4 einer inklusiven Grundschule ausgeprägt sind, kann nicht direkt in einem Satz beantwortet werden. Denn das Auftreten und somit auch die Ausprägung der Boden- und Deckeneffekte ist abhängig von dem jeweiligen Testformat bzw. von der im Testformat enthaltenen Anzahl an Items. Da das Auftreten von Boden- und Deckeneffekten allerdings definitiv bestätigt werden kann, kann somit auch die erste Hypothese (*Die Lernplattform Levumi weist innerhalb der Lerndomäne „Lesen“ in den Jahrgangsstufen 3/4 einer inklusiven Grundschule Boden- und Deckeneffekte auf.)* verifiziert werden. Auch Hypothese 2, dass die Häufigkeit von Deckeneffekten in den Schülerverläufen mit fortschreitenden Messzeitpunkten zunimmt, während die Häufigkeit des Auftretens eines Bodeneffekts gleichzeitig abnimmt, kann mit Blick auf Tabelle 12 (S. 78) als logische Schlussfolgerung eindeutig verifiziert werden. Zusätzlich kann hinzugefügt werden, dass Deckeneffekte vermehrt in der vierten Jahrgangsstufe auftreten, während Bodeneffekte häufiger in der Jahrgangsstufe 3 auftreten. Die dritte Hypothese - *die Lesetests zum Wortlesen und zum sinnentnehmenden Lesen weisen häufiger Deckeneffekte auf, als die Tests zum Silben- und Pseudowörterlesen* - kann ebenfalls als zutreffend eingeschätzt und mit der geringeren Itemanzahl begründet werden. Um die Fragestellung differenzierter beantworten zu können, wird im Folgenden noch einmal auf alle vier Lesetests einzeln eingegangen.

Wie die vorangegangene Datenauswertung beweisen konnte, treten in den Levumitests zur Lesekompetenz sowohl Boden- als auch Deckeneffekte auf. Als einzige Ausnahme treten in dem Leseflüssigkeitstest zum Silbenlesen weder Boden- noch Deckeneffekte auf. Dies lässt sich zum einen damit erklären, dass der Test zum Silbenlesen als einfachstes Testformat gilt und somit das Auftreten eines Bodeneffektes unwahrscheinlich wird (vgl. Jungjohann et al. 2017, S. 4). Zum anderen ist der Silbentest mit einer gesamten Itemanzahl von 134 sehr umfangreich. Da es nahezu unmöglich ist in einem computergestützten Verfahren eine so hohe Anzahl an Items zu lesen, ist es ohnehin unwahrscheinlich, dass ein Deckeneffekt die Testwerte einschränkt. Aufgrund dieser immens hohen Itemanzahl, wie es für einen Speedtest typisch ist, bietet sich die Analyse der Itemschwierigkeiten nach dem von Jungjohann, Rütter und DeVries (2018c) empfohlenen Vorgehen für diese Zwecke eher weniger an (vgl. S. 10). Denn durch die hohe Anzahl an Items, die aufgrund der Zeitbeschränkung auf eine Minute, nicht gelöst werden können, die aber aufgrund der Auswertung nach dem Speedtest-Verfahren als falsch gewertet werden, würde die Itemschwierigkeit wahrscheinlich schlechter

ausfallen als sie tatsächlich wäre. Da die Homogenität der Testschwierigkeit aber ein Kriterium der CBM-Tests ist und sich die gewählten Silben durch den KLA begründen lassen, wird von gleichschweren sowie angemessen schweren Items ausgegangen.

Tabelle 12 im vorherigen Kapitel zeigt, dass im Leseflüssigkeitstest zum Wörterlesen ebenfalls kein Bodeneffekt auftritt. Im Gegensatz zum Silbenlesen werden die Testwerte im Wörterlesen aber durch einen Deckeneffekt gestutzt. Das Auftreten der Deckeneffekte ist in der Jahrgangsstufe 4 mit ca. 13,64% aller Testwerte stärker ausgeprägt als in der Jahrgangsstufe 3, in der lediglich 6,04% der Testwerte durch einen Deckeneffekt eingeschränkt werden. Das Auftreten des Deckeneffekts in diesem Testformat kann durch die geringe Itemanzahl in Höhe von 61 Items erklärt werden. So ist es nicht verwunderlich, dass in der Jahrgangsstufe 3 vier Testergebnisse (≈ 2,2%) und in der Jahrgangsstufe 4 sieben Testergebnisse (≈ 4,55%) die maximale Itemanzahl von 61 korrekt gelesenen Items erreichen. Da im Unterschied zu den Silben- und Pseudowörterlesetests die Itemanzahl im Wortlesen relativ gering gehalten wurde, was bedeutet, dass die meisten SuS die meisten Items gelesen haben, bietet sich die Ermittlung des Schwierigkeitsindizes P_i innerhalb dieses Testformats mehr an.

Der durchschnittliche Schwierigkeitsindex in der Jahrgangsstufe 3 liegt bei $P_i \approx$ 55,07. Nach Döring und Bortz (2016) sind Items mit P_i = 80 - 100 als extrem leicht, Items mit P_i = 0 - 20 als extrem schwierig einzuschätzen. Sie empfehlen daher Items mit einem Schwierigkeitsindex von P_i = 20 - 80. Hinsichtlich dieser Ergebnisse ist das Auftreten der Deckeneffekte nicht mit der Itemschwierigkeit zu erklären, denn ein Schwierigkeitsindex von $P_i \approx$ 55,07 liegt nach Döring und Bortz (2016) im optimalen Bereich (vgl. S. 476), sondern muss, wie oben bereits vermutet, auf die geringere Itemanzahl zurückgeführt werden. Der durchschnittliche Schwierigkeitsindex in der Jahrgangsstufe 4 liegt bei $P_i \approx$ 73,69 und liegt somit auch im erwünschten Schwierigkeitsbereich. Der unterschiedliche Schwierigkeitsindizes der dritten und vierten Jahrgangsstufe lässt sich, wie schon im Auswertungskapitel beschrieben, damit erklären, dass die SuS der vierten Jahrgangsstufe insgesamt mehr Items korrekt beantworten konnten, sodass aufgrund der Speedtest-Variante der Schwierigkeitsindex in der Jahrgangsstufe 4 höher ausfällt als in der Jahrgangsstufe 3.

Sowohl in der dritten als auch in der vierten Jahrgangsstufe sind die beiden Items „*falsche*“ und „*Furcht*“ konstant schwieriger ausgefallen. Während der Schwierigkeitsindex des Items „*falsche*“ in der Jahrgangsstufe 3 bei $P_i \approx$ 25,1 und in der Jahrgangsstufe 4 bei $P_i \approx$ 51,96 liegt, erreicht das Item „*Furcht*“ in der dritten Jahrgangsstufe lediglich einen Schwierigkeitsindex von $P_i \approx$ 18,67 und von $P_i \approx$ 32,67 in der vierten Jahrgangsstufe. In der Jahrgangsstufe 4 erwies sich außerdem das Item „*Puls“* mit einem durchschnittlichen Index von

$P_i \approx 57{,}07$ als schwieriger im Vergleich zu anderen Items. Eine mögliche Erklärung für das schlechtere Abschneiden dieser Items ist, dass der Wörterlesetest neben den drei aufgezählten Items auch sehr ähnlich klingende Items enthält. So unterscheiden sich die Items „*plus*“ (Jgst. 3: $P_i \approx 64{,}88$; Jgst. 4: $P_i \approx 79{,}23$), „*Flasche*“ “ (Jgst. 3: $P_i \approx 66{,}11$; Jgst. 4: $P_i \approx 80{,}42$) und „*Frucht*“ (Jgst. 3: $P_i \approx 33{,}47$; Jgst. 4: $P_i \approx 61{,}25$) lediglich durch die Vertauschung zweier Grapheme voneinander. Eine Begründung, warum die letztgenannten Items besser abschneiden, könnte darin liegen, dass diese bereits öfter gelesen wurden oder im Fall der Items „*plus*“ und „*Puls*“, dass das Wort „*plus*“ möglicherweise näher am Alltag der SuS ist.

Werden die in dieser Studie erhaltenen Ergebnisse mit bereits bestehenden Erkenntnissen in der Literatur verglichen, werden sowohl Parallelen als auch Ungereimtheiten entdeckt. Eine Parallele stellt das Auftreten des Deckeneffekts im Wortlesen mit den Ergebnissen des systematischen Reviews von Jungjohann et al. (2018) dar. Denn dort wurde bereits auf das gehäufte Auftreten von Deckeneffekten in Lernverlaufsverfahren von älteren SuS hingewiesen, bei denen die *oral reading fluency*, d.h. die Anzahl korrekt gelesener Wörter innerhalb einer Minute, gemessen wurde (vgl. S. 112f.)

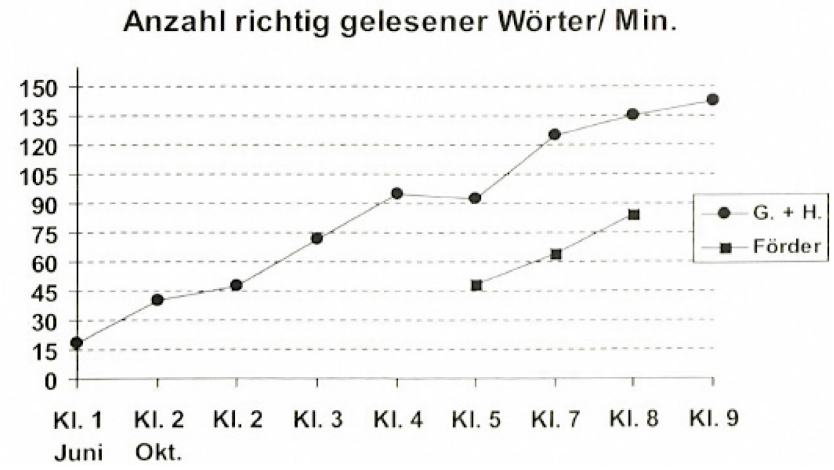


Abbildung 30: *Anzahl richtig gelesener Wörter pro Minute nach Walter (2008) (entnommen aus: Walter 2008, S. 75)*

Beim Vergleich der in dieser Studie erhaltenen Ergebnisse im Wörterlesetest mit den von Walter (2008) als Normwerte für den Wortlesetest seiner Lernfortschrittsdiagnostik Lesen mit einer Zeitbeschränkung auf eine Minute vorgestellten Ergebnisse (s. Abbildung 30), fallen schnell Ungereimtheiten auf. Denn beim Vergleich der durchschnittlich gelesenen Anzahl an Wörtern innerhalb des Levumitests in der dritten und vierten Jahrgangsstufe mit den in Abbildung 30 präsentierten Ergebnissen, wird direkt offensichtlich, dass die in dieser Studie erhaltenen Ergebnisse bei weitem nicht so hoch ausfallen wie bei Walter (vgl. S. 75). Erreichen die SuS bei Walter in der dritten Klasse bereits durchschnittlich 75 Wörter pro Minute, ist der

Levumi-Test im Wörterlesen auf eine gesamte Anzahl von 61 Testitems begrenzt. Ähnlich fällt auch der Vergleich mit den von Müller et al. (2013) gemessenen Werten aus. Denn innerhalb dessen erreichten SuS der Jahrgangsstufe 4 in einer CBM Testung zur Leseflüssigkeit Summenscores von über 40 gelesenen Silben und Wörtern innerhalb von nur 20 Sekunden (vgl. S. 138). Würde diese Testwert auf eine Minute hochgerechnet werden, würden diese sogar Walters (2008) Normwerte übersteigen. Zu beachten ist aber, dass in dem Müller et al. (2013) durchgeführten Testverfahren Wörter und Silben gemischt auftreten.

Diese immensen Unterschiede begründen sich nicht damit, dass die SuS in der Levumi-Testung schlechtere Ergebnisse erzielten, sondern in der Konzeption des Testformates. Denn während bei Walter (2008) oder Müller et al. (2013) der Antwortweg lediglich über den Probanden und den Rater läuft, ist bei den Levumi-Tests zusätzlich der Computer dazwischen geschaltet, sodass die SuS der 3. Jahrgangsstufe bei Levumi auf eine durchschnittliche Anzahl von ca. 34 richtig gelesen Wörtern pro Minute kommen. Die SuS der Jahrgangsstufe 4 erreichen hingegen schon eine Anzahl von etwa 45 Wörtern pro Minute, im Vergleich zu ca. 95 richtig gelesenen Wörtern bei Walter (vgl. 2008, S. 75). Normwerte eines computergestützten Verfahrens, welches der Levumi-Plattform ähnelt, konnten zum Vergleich nicht gefunden werden. An dieser Stelle sollte jedoch außerdem erwähnt werden, dass einige der SuS, die mit der Online-Plattform Levumi getestet wurden, bessere Summenscores hätten erreichen können, wäre im Test eine höhere Itemanzahl vorhanden. Denn einige SuS erhielten die Rückmeldung, dass der Test beendet ist, da die maximale Itemanzahl bearbeitet wurde und eben nicht, dass die Testzeit abgelaufen ist. Wären mehr Items zur Verfügung gewesen, hätten die SuS aufgrund der verbliebenen Zeit möglicherweise einen höheren Summenscore erreichen können.

Auch beim Vergleich der hier erhalten Schülerleistungen mit dem im zweiten Kapitel dargestellten theoretischen Hintergrund fällt auf, dass die SuS mit einem Maximalwert von 61 Items relativ weit entfernt sind von dem von Scheerer-Neumann (2015) angegeben Wert von 100 bis 150 gelesenen Wörtern pro Minute, welcher als Kennzeichen des automatisierten Leseprozesses gilt (vgl. S. 73). Ebenso muss aber auch hier darauf hingewiesen werden, dass aufgrund des Testformats bzw. des Lösungswegs über den Schüler, den Testdurchführer sowie den Computer mehr Zeit in Anspruch genommen wird, als direkte Testverfahren oder als beim verinnerlichten Leseprozess. In Bezug zu den in Kapitel 2.2 vorgestellten Entwicklungstheorien zur Lesekompetenz kann festgehalten werden, dass die Testwerte der Leseflüssigkeitsmessungen Aussagen über die unterschiedlichen Verarbeitungswege des Wortlesens ermöglichen. Da die SuS während einer Minute deutlich mehr sinnvolle Wörter lesen können als Pseudowörter, kann die höhere Effizienz des direkten Leseweges, d. h.,

dass direkte Erkennen eines Wortes aufgrund phonologischer oder visueller Auffälligkeiten, durch die hier erhaltenen Testergebnisse bestätigt werden (vgl. Gebhardt et al. 2016a, S. 6). Da die SuS in der Lage sind die sinnfrei erschaffenen sowie die sinnvollen Wörter zu erlesen, sind sowohl der lexikalische als auch der nichtlexikalische Leseweg bei den SuS dieser Studie bereits ausgebildet.

Wie Tabelle 12 (s. S. 78) ebenfalls zu entnehmen ist, tritt im Pseudowörtertest sowohl in der dritten als auch in der vierten Jahrgangsstufe ein Bodeneffekt auf. Da die Ausprägung dessen während des gesamten Schuljahres, in dem die Testungen durchgeführt wurden, abnehmen, sodass zu Messzeitpunkt 4 in der Jahrgangsstufe 4 gar kein Testergebnis mehr unter die Grenze des Bodeneffekts fällt, kann davon ausgegangen werden, dass die SuS einen Lernfortschritt vollziehen. Da der Bodeneffekt besonders in der Jahrgangsstufe 3 stark auftritt, da während allen Testungen durchschnittlich ca. 9,5% der Summenscores durch den Bodeneffekt begrenzt werden und während der ersten Messzeitpunktes sogar ein Fünftel der Testergebnisse einen Bodeneffekt aufweisen, kann überlegt werden, ob sich hier die Messung auf einer niedrigeren Niveaustufe eignet. Da die Niveaustufen N3a und N3b besonders selten genutzte Grapheme, wie z. B. <j, v, ß, x, y, ö>, enthält, sollte eher die Nutzung der Niveaustufen N2a und N2b in Betracht gezogen werden (vgl. Jungjohann et al. 2018b, S. 4; Abbildung 5, S. 44). Abgesehen von dieser Option das Auftreten des Bodeneffekts zu vermeiden, ist die Stärke des Bodeneffekts auch weniger stark einzuschätzen, wenn beachtet wird, dass die SuS, um unter den Bodeneffekt zu fallen, immerhin noch neun Items richtig lösen können. Der Nachteil des Auftretens von Boden- und Deckeneffekten liegt darin, dass die Differenzierungskraft der Testergebnisse abnimmt. Da die SuS in dieser Studie immerhin neun Items richtig lösen können, aber dennoch unter der Grenze des Bodeneffekts liegen, können trotzdem differenzierte Aussagen über die Schülerleistungen getroffen werden. Schaut man sich die erhaltenen Daten an, wird offensichtlich, dass der geringste Summenscore während allen Messzeitpunkten und beider Jahrgangsstufen bei 5 liegt und somit bei genauerer Betrachtung dieses Summenscores auch weitere differenzierte Aussagen möglich sind.

Ebenso wie bei dem Silbentest bietet sich auch im Pseudowortlesen die Schwierigkeitsanalyse aufgrund der bereits genannten Punkte nicht an, da die Anzahl der Items mit 189 ebenfalls sehr groß ausfällt.

Wird noch einmal Tabelle 12 betrachtet, fällt auf, dass im sinnentnehmenden Lesetest lediglich in der dritten Jahrgangsstufe Bodeneffekte auftreten. Dies lässt sich jedoch ganz einfach mit der sehr hoch gesetzten Grenze von 25% der Items begründen. Wie bereits erwähnt, resultiert dies daraus, da die Wahrscheinlichkeit das richtige Item aus Zufall auszuwählen

bereits bei 25% liegt. Bodeneffekte sind nicht wünschenswert, da sie in den unteren Merkmalsbereichen keine Differenzierungskraft haben. Da die Grenze des Bodeneffekts bei 25% angesetzt wurde, bedeutet dies, dass 15 oder weniger Items korrekt gelöst werden müssen, um unter diese Grenze zu fallen. Daher kann davon ausgegangen werden, dass Summenscores, aufgrund der hohen Itemanzahl, auch unter dieser Grenze noch differenziert werden können. Da die Ausprägung der Bodeneffekte von anfänglich 9 Probanden (≈ 20,45%) außerdem auf einen einzigen Probanden während des letzten Messzeitpunktes langsam absinkt, ist der auftretende Bodeneffekt in dem sinnentnehmenden Lesetest in der dritten Jahrgangsstufe als weniger schwerwiegend einzuschätzen. Sollten dennoch Bodeneffekte auftreten, aufgrund dessen keine differenzierten Aussagen über den Leistungsstand der SuS mehr möglich sind, so haben Lehrkräfte mit der Online-Plattform Levumi die Möglichkeit die Niveaustufe zu verändern. Diese Studie wurde auf der höchsten Niveaustufe durchgeführt, sodass zur Differenzierung in die unteren Leistungsbereiche noch Spielraum besteht (s. Abbildung 5, S. 44).

Anders gestaltet sich dies hinsichtlich des aufgetretenen Deckeneffekts in dem Leseverständnistest der Levumi-Plattform. Denn zum einen tritt dieser in der Jahrgangsstufe 3 und der Jahrgangsstufe 4 auf. Zum anderen beeinflusst der Deckeneffekt besonders in der Jahrgangsstufe 4 die Testergebnisse in einem sehr starken Ausmaß. Denn insgesamt werden während aller Testzeitpunkte ca. 31,13% der Testergebnisse dieser Jahrgangsstufe durch den Deckeneffekt eingeschränkt, zu Messzeitpunkt 4 waren es sogar 55%. Werden auch diese Testergebnisse mit bereits ermittelten Normwerten von Walter (2013) verglichen, muss zuerst darauf hingewiesen werden, dass die Tests der VSL - Verlaufsdiagnostik sinnerfassendes Lesen nicht auf Satzebene, wie der Levumi-Test, sondern auf Textebene stattfindet. Darüber hinaus haben die SuS für die von Walter konzipierten Tests lediglich 4 Minuten und nicht 7 Minuten wie auf der Levumi-Plattform. Maximal lösten die SuS am Anfang der dritten Klassen in der VSL 24 Items korrekt, gegenüber 61 in Levumi. In der Schuljahresmitte lösten mit der VSL 50% der SuS schlechtere und 50% bessere Ergebnisse als der Rohwert von 13. Dementgegen lag der Median der hier ermittelten Daten zu Messzeitpunkt 2 und 3 bei Md ≈ 32,5 bzw. Md≈ 38,5. Auch beim Vergleich der Ergebnisse der vierten Klasse fallen die Ergebnisse bei Walter (2013) schlechter aus. Denn so liegt der Median des zweiten und dritten Messzeitpunktes in der Levumi-Testung bei Md≈ 47 bzw. Md≈ 56,5, im Gegensatz zu 50% der SuS, die einen höheren bzw. einen niedrigen Wert als einen Rohwert von 19 richtigen Items erhalten (vgl. Walter 2013, S.86f.). Wird berücksichtigt, dass die Probanden in der VSL beinahe die Hälfte der Zeit zur Verfügung haben, aber dafür einen Text lesen müssen, in denen viele Sätze länger ausfallen als im sinnentnehmenden Lesetest von Levumi und

außerdem innerhalb eines Satzes auch mehrere Wörter fehlen können, ist das schlechtere Abschneiden der Ergebnisse der VSL nicht überraschend. Darüber hinaus kann auch das Textverständnis als schwieriger als das Verständnis einzelner voneinander unabhängiger Sätze eingeschätzt werden. Denn für das Verständnis eines Textes müssen SuS aktiv auf das Wissen bereits gelesener Textpassagen zurückgreifen können und gleichzeitig diesen Wissenstand weiter ausbauen (vgl. Bredel et al. 2011, S. 74).

Genau wie bei dem Leseflüssigkeitstest zum Wörterlesen bietet sich aufgrund der geringeren Itemanzahl eine Schwierigkeitsanalyse der Items hier an. Denn „die Itemschwierigkeiten beeinflussen […] ganz wesentlich die Verteilung der Testwerte“ (Döring & Bortz 2016, S. 476). Im vorherigen Kapitel zur Datenauswertung wurden bereits die Itemschwierigkeiten der Items des sinnentnehmenden Lesetests dargelegt. Die durchschnittliche Itemschwierigkeit schwankt zwischen den beiden Jahrgangsstufen zwischen $P_i \approx 57{,}26$ in der dritten Jahrgangsstufe und $P_i \approx 78{,}78$ in der vierten Jahrgangsstufe. Es wurde bereits auf das schlechtere Abschneiden der Items „*über*“, „*weder*“ und insbesondere des Items „*In*“ eingegangen (s. Tabelle 3, S. 58). Bei der Betrachtung der weiteren Items ist darüber hinaus auch auffällig, dass Nomen, Verben und Adjektive häufiger richtig gelöst werden als andere Wortarten. Dabei sorgten besonders Junktoren und Präpositionen, wie beispielsweise „*mit*“, „*Durch*“, „*weil*“, „*Wenn*“, „*für*“ oder „*außer*“, für Schwierigkeiten. Dies gilt sowohl für die Jahrgangstufe 3, als auch für die vierte Jahrgangsstufe. Eine Ausnahme lässt sich jedoch in der dritten Jahrgangsstufe zu Messzeitpunkt 3 feststellen. Denn in diesem erreichen die Items „*unter*“ ($P_i \approx 81{,}82$), „*aber*“ ($P_i \approx 77{,}27$), „*bevor*“ ($P_i = 75$), „*Wenn*“ ($P_i \approx 72{,}73$), „*Während*“ ($P_i \approx 72{,}73$) etc. die höchsten Schwierigkeitsindizes, was bedeutet, dass diese Items den SuS zu dem genannten Testzeitpunkt am leichtesten fielen (s. Anhang, S. XXX). Eine Ursache dafür könnte darin liegen, dass vor der Durchführung der dritten Messung Präpositionen und Junktoren unterrichtlich bearbeitet wurden. Ob tatsächlich unterrichtliche Instruktionen bezüglich dieses Themenbereiches vorgenommen wurden, ist jedoch nicht bekannt.

Abgesehen von den ausgeprägten Deckeneffekten bezüglich der hohen Summenscores der SuS, ist während der Testdurchführung außerdem offensichtlich geworden, dass besonders SuS der vierten Jahrgangsstufe den Test vor den eigentlich vorgesehenen sieben Minuten Testzeit abschließen konnten. So konnten einige SuS den Test bereits nach einer Zeit von 3,5 Minuten beenden. In diesem Fall kann man die starke Ausprägung im Gegensatz zum Leseflüssigkeitstest Wörterlesen nicht nur durch einen eventuell zu leichten Schwierigkeitsgrad sowie eine zu geringe Anzahl an Items erklären, sondern außerdem mit einer mit 7 Minuten wahrscheinlich zu lang eingeschätzten Zeitbegrenzung. Um das Auftreten eines

Deckeneffekts zu vermeiden, kann meist auf eine höhere Niveaustufe zurückgegriffen werden. In der hier durchgeführten Langzeitstudie wurde jedoch bereits auf der höchsten Niveaustufe getestet. Um das starke Auftreten der Deckeneffekte zu verringern, sollte daher eine Testanpassung erfolgen. Eine Möglichkeit kann darin bestehen, dass der Testzeitraum von 7 Minuten auf drei bis vier Minuten gekürzt wird oder dass dem Test mehr Items hinzugefügt werden. Der Deckeneffekt zeigt besonders zu Messzeitpunkt 4 in der vierten Jahrgangsstufe eine starke Ausprägung, da in dieser Testung mehr als die Hälfte (≈ 55%) der Testwerte durch den Deckeneffekt eingeschränkt werden. Da das Diagnoseinventar Levumi bisher ausschließlich für die Anwendung in der Grundschule konzipiert wurde, ist die starke Ausprägung der Deckeneffekte am Ende der Jahrgangsstufe 4 jedoch als weniger schwerwiegend einzuschätzen als es auf den ersten Blick erscheint. Denn somit zeigt der Test, dass er tatsächlich lediglich zur Anwendung in der Grundschule geeignet ist und Viertklässler zumindest im sinnentnehmenden Lesetest die Erwartungen übersteigen. Nichtsdestotrotz arbeitet das Forscherteam auch an einer Testerweiterung zur Implementierung von Levumi in die Sekundarstufe (vgl. Jungjohann et al. 2018b, S. 7).

Insgesamt betrachtet, kann das Auftreten der Bodeneffekte als problematischer eingeschätzt werden als das umfangreichere Auftreten der Deckeneffekte. Denn der Anspruch der Lernverlaufsdiagnostik liegt darin insbesondere SuS mit Lernschwierigkeiten eine bessere Förderung zu ermöglichen (vgl. Diehl 2011, S. 166). Bezüglich dessen ist auch das Kriterium der Testfairness von besonderer Relevanz. Testfairness meint, dass die Tests beispielsweise SuS mit Migrationshintergrund oder mit zugeschriebenen Förderbedarfen nicht benachteiligen dürfen (vgl. Moosbrugger & Kelava 2008, S. 23). Aufgrund der Lernschwierigkeiten kann geschlussfolgert werden, dass diese SuS mit höherer Wahrscheinlichkeit unter der Grenze des Bodeneffekts liegen. Sollte dies nun der Fall sein und keine zuverlässigen Aussagen über den Lernverlauf der SuS möglich sein, verfehlt das Instrument zur Lernverlaufsdiagnostik seinen eigentlichen Anspruch. Im Gegensatz dazu ist das Auftreten von Deckeneffekten beinahe als wünschenswert zu bezeichnen. Denn Deckeneffekte sprechen dafür, dass die SuS bereits über einige Kompetenzen verfügen und scheinbar einen besseren Kenntnisstand oder einen höheren Kompetenzstand aufweisen als von den Testentwicklern oder Lehrkräften eingeschätzt.

Deckeneffekte sind auch typisch für die Intelligenztestung. Ähnlich zu den hier erhaltenen Ergebnissen sind Intelligenztests meist gut geeignet, um im mittleren Leistungsbereich zu messen, sind aber weniger geeignet, um sichere Aussagen über die höheren Leistungsbe-

reiche zu machen. So erfassen die meisten Intelligenztests den IQ bis hin zu einem bestimmten Grenzwert (meist ein IQ von 130) sehr genau, was völlig ausreichend ist, um beispielsweise Selektionsentscheidungen in der Praxis treffen zu können. Sollte jedoch der genaue IQ-Wert von Bedeutung sein, sodass Deckeneffekte vermieden werden müssen, so werden weitere spezifische Tests durchgeführt, indem zum Beispiel auf Tests für ältere SuS zurückgegriffen wird. Die einzige Schwierigkeit bei diesem sogenannten *above-level-testing* besteht dann darin, dass für die jüngeren Probanden keine Normwerte vorliegen (vgl. Vock, Preckel & Holling 2007, S. 129f.). Holling, Preckel und Vock (2004) weisen außerdem daraufhin, dass sich im Falle des Auftretens von Boden- und Deckeneffekten auch spezifische Tests, beispielsweise zur Ermittlung der Intelligenz bei SuS mit Lernbehinderung oder mit Hochbegabung, anbieten (vgl. S. 60).

Ähnlich zu dem in der Intelligenztestung vorgenommenen Verfahren kann auch hier vorgegangen werden. Denn zum einen sprechen im Falle von Deckeneffekten die hohen Summenscores bereits für sich und gegen eine notwendige Optimierung der Fördermaßnahme. Sollte aber auch hier der genaue Summenscore interessant sein, so können auch hier Testverfahren einer höheren Niveaustufe oder einer höheren Jahrgangsstufe, im Sinne des *above-level-testings*, verwendet werden, um genauere Ergebnisse zu erhalten. Gleiches ist selbstverständlich auch mit Testverfahren einer niedrigen Klassenstufe denkbar. Dementgegen steht jedoch die Testdurchführung auf der Niveaustufe N4. Denn während bei einem aufgetretenen Bodeneffekt in diesem Fall auf eine niedrigere Niveaustufe zurückgegriffen werden kann, um differenziertere Ergebnisse zu erhalten, ist dies bei den Deckeneffekten nicht der Fall. N4 ist bisher die höchste Niveaustufe der Levumitests, sodass eine Steigerung bis heute noch nicht möglich ist. Da im Forschungsprojekt Levumi aber auch Tests für die Sekundarstufe in Planung sind, sollte dies bald keine Schwierigkeit mehr darstellen.

8 Zusammenfassung und Ausblick

Die vorliegende Arbeit thematisiert die Grenzen des Testens in Lernverlaufsmessungen. Genauer gesagt wurden mit dem Lernverlaufsmonitoring Levumi erhobene Daten auf das Auftreten von Boden- und Deckeneffekten untersucht. Folglich war das Ziel der Arbeit herauszustellen, wie stark die Ausprägung der Boden- und Deckeneffekte in der Lerndomäne Lesen des Diagnoseinventars Levumi vorliegt.

Um das Ziel zu erreichen, wurden die in dieser Arbeit ausgewerteten Daten in den Jahrgangsstufen 3 und 4 mittels einer Langzeitstudie im Panel-Design in einer inklusiven Grundschule in NRW erhoben. Da die Lerndomäne des Lesens in der Plattform Levumi einen Test zum sinnentnehmenden Lesen und drei zur Leseflüssigkeit enthält, musste jeder der 87 Probanden vier Tests durchführen. Infolge der Panel-Mortalität sowie Schwierigkeiten mit der Datenspeicherung, schwankt die Anzahl der Testteilnehmer jedoch während allen Testungen. Um Boden- und Deckeneffekte ermitteln zu können, wurden anhand bestimmter Kriterien Grenzen gesetzt, ab wann von einem Decken- bzw. bis wann von einem Bodeneffekt gesprochen werden kann. Die Grenze des Deckeneffekts wurde bei 95% angesetzt, d.h. die SuS müssen mindestens 95% der Items richtig lösen, um von einem Deckeneffekt sprechen zu können. Die Grenze des Bodeneffekts wurde, mit Ausnahme des sinnentnehmenden Lesetests, bei 5% angesiedelt. Folglich weisen die Testergebnisse einen Bodeneffekt auf, wenn lediglich 5% der Items oder weniger richtig gelöst wurden. Aufgrund der unterschiedlichen Itemanzahlen der einzelnen Tests bedeutet die Prozentgrenze demnach auch eine unterschiedliche Anzahl an Items, die gelöst werden können, um unter oder über die genannten Grenzen zu fallen. Da im sinnentnehmenden Lesetest die Zufallswahrscheinlichkeit für die Auswahl des richtigen Items bereits bei 25% liegt, wurde dies auch als Grenze des Bodeneffekts innerhalb dieses Testformats ausgewählt.

Infolge dieser Analyse konnte herausgefunden werden, dass lediglich in den Tests zum sinnentnehmenden Leseverständnis und zum Wörterlesen Deckeneffekte auftreten. Dabei treten diese besonders stark in den Ergebnissen des Leseverständnistests der Jahrgangsstufe 4 zutage. Denn dort erreichten 55% der SuS während des vierten Messzeitpunktes einen Summenscore von mind. 58 richtig gelösten Items. Insgesamt weisen in der Jahrgangsstufe 4 während allen Messzeitpunkten durchschnittlich 31,13% der Testergebnisse einen Deckeneffekt auf, in der Jahrgangsstufe 3 hingegen lediglich 9,66%. Im Test zum Wörter lesen fällt die Ausprägung des Deckeneffekts noch geringer aus, hier fallen ca. 6,04% der Testergebnisse der Jahrgangsstufe 3 und 13,64% der Jahrgangsstufe 4 über die gesetzte Grenze.

Die Ergebnisse des sinnentnehmenden Lesetests der Jahrgangsstufe 3 weisen als einziges sowohl einen Boden- als auch einen Deckeneffekt auf. Dies liegt jedoch an der mit 25% sehr hoch gesetzten Bodeneffektsgrenze unter die immerhin 10,8% der Testwerte fallen. Wie oben bereits beschrieben, muss diese Grenze jedoch so hoch ausfallen, da anzunehmen ist, dass SuS, welche weniger als 25% der Items richtig lösen, lediglich Ratestrategien anwenden und aus Zufall das korrekte Item auswählen.

So treten Bodeneffekte hauptsächlich im Pseudowörterlesen und insbesondere in der Jahrgangsstufe 3 auf, in der 9,5% der Testergebnisse durch einen Bodeneffekt begrenzt werden. In der Jahrgangsstufe 4 sind es hingegen nur noch ca. 1,95%. Aufgrund der hohen Itemanzahl und der Komplexität des Pseudowörterlesen ist das Auftreten dieser jedoch nicht überraschend. Während die Ausprägung der Deckeneffekte in dem Wörterlesetest oder im Leseverständnistest mit den Messzeitpunkten stetig zu nimmt, nimmt die Ausprägung des Bodeneffektes ab, sodass am Ende der Langzeitstudie, d.h. zu Messzeitpunkt 4, in der Jahrgangsstufe 4 beispielsweise kein Bodeneffekt in dem Test zum Pseudowörterlesen mehr zu vernehmen ist.

Um den Silbentest nicht außen vorzulassen, sei an dieser Stelle noch einmal erwähnt, dass dieser, laut der hier stattgefundenen Analyse, frei von Boden- und Deckeneffekten ist.

Bereits im vorherigen Kapitel wurden Empfehlungen zum weiteren Umgang mit den hier erhaltenen Daten genannt. So wurde bezüglich der Reduktion des Deckeneffekts vorgeschlagen, die Itemanzahl im sinnentnehmenden Lesetest sowie im Wörterlesetest zu erhöhen, oder auch eine höhere Niveaustufe zu etablieren. Darüber hinaus kann im Hinblick auf den Test zum Leseverständnis auch die Testzeit in Höhe von 7 Minuten, um die Hälfte verringert werden, da einige der Probanden bereits nach 3,5 Minuten den Test beenden konnten.

Eine weitere Möglichkeit, um das Auftreten dieser Grenzeffekte zu vermeiden, ist das adaptive Testen. In einem adaptiven Testsystem werden den Probanden nicht alle enthaltenen Items zur Lösung angeboten, sondern nur diejenigen, die dem Leistungsstand des Probanden entsprechen. Nach Angaben der Entwickler ist die Erweiterung des Diagnoseinstruments hin zu einem adaptiven Testsystem in der nächsten Zeit vorgesehen (vgl. Mühling et al. 2017, S. 560).

Insgesamt kann als Konsequenz für die Schulpraxis festgehalten werden, dass sich die Anwendung der Levumitests trotz aufgetretener Boden- und Deckeneffekte durchaus eignet. Da im Falle von Bodeneffekten eine geringere Niveaustufe gewählt werden kann, stellt das Auf-

treten dieser keine große Schwierigkeit dar. Bezüglich des Auftretens des Deckeneffekts bietet sich jedoch eine Testanpassung an, da nicht auf eine höhere Niveaustufe zurückgegriffen werden kann. Mögliche Optionen, um dennoch die Ausprägung des Deckeneffekts zu verringern, ist die Reduktion der Testzeit des sinnentnehmenden Lesetests oder die Bereitstellung einer größeren Itemanzahl. Die Reduktion der Testzeit im sinnentnehmenden Lesetest stellt dabei die wahrscheinlich einfachste, aber auch wichtigste Option dar. Denn der Deckeneffekt des Leseverständnistests sollte aufgrund der starken Ausprägung definitiv ausgebügelt werden.

Als Ausblick kann vermerkt werden, dass die Analyse, der in dieser Longitudinalstudie erhaltenen Daten, mit der Betrachtung der Boden- und Deckeneffekte keinesfalls ausgeschöpft ist. Zum einen kann die Analyse der Grenzen des Messens im Lesen mit der Online-Plattform Levumi weiter vertieft werden und neben der in dieser Studie fokussierten deskriptiven Statistik, auch ein Ausblick auf die Auswertung mittels Inferenzstatistik geworfen werden. Ebenso relevant kann die Überprüfung der hier vorliegenden Daten von Bedeutung sein; eine andere Schule, eine größere Stichprobe oder auch ein anderes Einzugsgebiet der Schule kann zu anderen Ergebnissen führen. Zum anderen ist besonders hinsichtlich der Lesekompetenz der Unterschied zwischen den Leistungen von Schülerinnen und Schülern interessant zu betrachten. Denn in der Entwicklung der Lesekompetenz zeigen sich häufig Geschlechtseffekte zugunsten der Mädchen (vgl. Schneider 2017, S. 95), sodass eine Untersuchung der vorliegenden Daten auf bestehende Geschlechtseffekte vorgenommen werden kann.

Auch eine Analyse der Testfairness mit besonderer Berücksichtigung von SuS mit Lernschwierigkeiten und/oder mit Migrationshintergrund kann aufschlussreiche Informationen über die Testkonstruktion liefern. Innerhalb dieser Studie weisen jedoch zu wenig der SuS einen Migrationshintergrund oder eine Lernschwierigkeit auf, um repräsentative Ergebnisse zu erhalten.

III Literaturverzeichnis

Ardoin, Scott P., Christ, Theodore J., Morena, Laura S., Cormier, Damien C. & Klingbeil, David A. (2013). A systematic review and summarization of the recommendations and research surrounding Curriculum-Based Measurement of oral reading fluency (CBM-R) decision rules. *Journal of School Psychology*, 51, p. 1-18.

Balt, Miriam, Ehlert, Antje & Fritz, Annemarie (2017). Theoriegeleitete Testkonstruktion dargestellt am Beispiel einer Lernverlaufsdiagnostik für den mathematischen Anfangsunterricht. *Empirische Sonderpädagogik*, 2, S. 165-183.

Baumann, Monika (2006). Lesetests. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch* (Band II). 2. Auflage (S. 869-882). Paderborn: Ferdinand Schöningh.

Bertschi-Kaufmann, Andrea (2010). Einsichten in Leseverhalten und Lesekönnen. In G. Schulz (Hrsg.), *Lesen lernen in der Grundschule. Lesekompetenz und Leseverstehen. Förderung von Bücherwelten* (S. 24-36). Berlin: Cornelsen Verlag Scriptor.

Bertschi-Kaufmann, Andrea (2011). Lesekompetenz – Leseleistung – Leseförderung. In A. Bertschi-Kaufmann (Hrsg.), *Lesekompetenz Leseleistung Leseförderung*. Grundlagen, Modelle und Materialien (4.Auflage, S. 8-17). Seelze: Kallmeyer in Verbindung mit Klett.

Blumenthal, Yvonne, Kuhlmann, Kristin & Hartke, Bodo (2014). Diagnostik und Prävention von Lernschwierigkeiten im Aptitude Treatment Interaction-(ATI-) und Response to Intervention-(RTI-)Ansatz. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (Tests und Trends, Band 12) (S. 61-81). Göttingen: Hogrefe Verlag.

Braun, Dorothee (2010). Leseschwierigkeiten erkennen, Schüler individuell fördern. In G. Schulz (Hrsg.), *Lesen lernen in der Grundschule. Lesekompetenz und Leseverstehen. Förderung von Bücherwelten* (S. 174-183). Berlin: Cornelsen Verlag Scriptor.

Bredel, Ursula, Fuhrhop, Nanna & Noack, Christina (2011). *Wie Kinder lesen und schreiben lernen*. Tübingen: Narr Francke Attempto Verlag.

Bürgermeister, Anika, Klieme, Eckhard, Rakoczy, Katrin, Harks, Birgit & Blum, Werner (2014). Formative Leistungsbeurteilung im Unterricht: Konzepte, Praxisberichte und ein neues Diagnoseinstrument für das Fach Mathematik. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (Tests und Trends, Band 12) (S. 41-60). Göttingen: Hogrefe Verlag.

Dehn, Mechthild (2010). Lesenlernen und Leseförderung. In G. Schulz (Hrsg.), *Lesen lernen in der Grundschule. Lesekompetenz und Leseverstehen. Förderung von Bücherwelten* (S. 136-150). Berlin: Cornelsen Verlag Scriptor.

Deno, Stanley L., Mirkin, Phyllis K. & Chiang, Bertram (1982). Identifying Valid Measures of Reading. *Exceptional Children,*1, p. 36-45.

Deno, Stanley L. (1985). Curriculum-Based Measurement: The Emerging Alternative. *Exceptional Children,* 3, p. 219-232.

Deno, Stanley L. (2003). Developments in Curriculum-Based Measurements. *The Journal of Special Education,* 3, p. 184-192.

Deutsches Institut für Menschenrechte (2006). *Übereinkommen über die Rechte von Menschen mit Behinderungen vom 13. Dezember 2006*. Online verfügbar unter: https://www.institut-fuer-menschenrechte.de/fileadmin/user_upload/PDF-Dateien/Pakte_Konventionen/CRPD_behindertenrechtskonvention/crpd_b_de.pdf [letzter Zugriff: 20.06.2018]

Diehl, Kerstin, Hartke, Bodo & Knopp, Eva (2009). Curriculum-Based Measurement & Leerlingonderwijsvolgsysteem – Konzepte zur theoriegeleiteten Lernfortschrittsmessung im Anfangsunterricht Deutsch und Mathematik? *Zeitschrift für Heilpädagogik*, 4, S. 122-130.

Diehl, Kerstin (2010). Lesen und Schreiben lernen. In B. Hartke, K. Koch & K. Diehl (Hrsg.), *Förderung in der schulischen Eingangsstufe* (S. 55-90). Stuttgart: W. Kohlhammer.

Diehl, Kerstin (2011). Innovative Lesediagnostik – Ein Schlüssel zur Prävention von Lese-Rechtschreibschwierigkeiten. *Zeitschrift für Heilpädagogik,* 5, S. 164-172.

Döring, Nicola & Bortz, Jürgen (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Auflage). Heidelberg: Springer Verlag.

Dummer-Smoch, Lisa & Hackethal, Renate (2007). *Kieler Leseaufbau. Handbuch* (7. Auflage). Kiel: Veris Verlag.

Eckermann, Johann Peter (1836). *Gespräche mit Goethe in den letzten Jahren seines Lebens* (Band 2). Leipzig: Brockhaus.

Ennemoser, Marc, Marx, Peter, Weber, Jutta & Schneider, Wolfgang (2012). Spezifische Vorläuferfähigkeiten der Lesegeschwindigkeit, des Leseverständnisses und des Rechtschreibens: Evidenz aus zwei Längsschnittstudien vom Kindergarten bis zur 4. Klasse. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 2,* S. 53-67.

Fink, Andreas, Luttenberger, Silke, Krammer, Andrea, Macher, Daniel, Papousek, Ilona, Weiss, Elisabeth M., Paechter, Manuela (2015). Die Veränderung kognitiver Fähigkeiten über die Sommerferien. *Psychologie in Erziehung und Unterricht*, 62, S. 303-315.

Forster, Maria (2005). Phonologische Bewusstheit als zentrale Voraussetzung für das Lesen: Möglichkeiten der Diagnose und Förderung. In E. Gläser & G. Franke-Zöllmer (Hrsg.), *Lesekompetenz fördern von Anfang an. Didaktische und methodische Anregungen zur Leseförderung* (S. 36-49). Baltmannsweiler: Schneider Verlag Hohengehren.

Förster, Natalie & Souvignier, Elmar (2011). Curriculum-Based Measurement: Developing a Computer-Based Assessment Instrument for Monitoring Student Reading Progress on Multiple Indicators. *Learning Disabilities: A Contemporary Journal*, 9, p. 21-44.

Förster, Natalie, Kuhn Jörg-Tobias & Souvignier, Elmar (2017). Kurzbeitrag: Normierung von Verfahren zur Lernverlaufsdiagnostik. *Empirische Sonderpädagogik*, 2, S. 116-122.

Frey, Hanno (2010). *Lesekompetenz verbessern? Lesestrategien und Bewusstmachungsverfahren.* Münster: Waxmann Verlag.

Frith, U. (1985). Beneath the Surface of Developmental Dyslexia. In K. Patterson, J. C. Marshall & M. Coltheart (Eds.), *Surface dyslexia. Neuropsychological and cognitive studies of phonological reading* (p. 301-330). London: Lawrence Erlbaum Associates.

Fuchs, Lynn S. (2004). The Past, Present, and Future of Curriculum-Based Measurement Research. *School Psychologiy Review*, 2, p. 188-192.

Galuschka, Katharina & Schulte-Körne, Gerd (2015). Evidenzbasierte Interventionsansätze und forschungsbasierte Programme zur Förderung der Leseleistung bei Kindern und Jugendlichen mit Lesestörung – Ein systematischer Review. *Zeitschrift für Erziehungswissenschaft*, 3, S. 473-487

Garbe, Christine (2010). Wie werden Kinder zu engagierten und kompetenten Lesern? In G. Schulz (Hrsg.), *Lesen lernen in der Grundschule. Lesekompetenz und Leseverstehen. Förderung von Bücherwelten* (S. 9-47). Berlin: Cornelsen Verlag Scriptor.

Gebhardt, Markus, Diehl, Kirsten & Mühling, Andreas (2016a). *Lern-Verlaufs-Monitoring LEVUMI Lehrerhandbuch*. Version 1.1. Online verfügbar unter: https://eldorado.tu-dortmund.de/bitstream/2003/35765/2/CBM_Lehrerhandbuch%20LEVUMI_final_1.1%20September.pdf [letzter Zugriff: 14.04.2018].

Gebhardt, M., Heine, J-H., Zeuch, N. & Förster, N. (2015). Lernverlaufsdiagnostik im Mathematikunterricht der zweiten Klasse. Raschanalysen zur Adaptation eines Testverfahrens für den Einsatz in inklusiven Klassen. Empirische Sonderpädagogik, (3), 206-222. Verfügbar unter: http://www.psychologie-aktuell.com/fileadmin/download/esp/3-2015_20150904/esp_3-2015_206-222.pdf

Gebhardt, Markus, Diehl, Kirsten & Mühling, Andreas (2016b). Online Lernverlaufsmessung für alle SchülerInnen in inklusiven Klassen. *Zeitschrift für Heilpädagogik, 66*, S. 444-454.

Gebhardt, Markus & Jungjohann, Jana (2018, im Druck). Digitale Unterstützung bei der Dokumentation von Verhaltens- und Leistungsbeurteilungen. In B. Meyer, T. Tretter, U. Englisch, (Hrsg.), *Praxisleitfaden auffällige Schüler und Schülerinnen.* Weinheim, Basel: Beltz.

Gehle-Davids, Swenja (2015). *Schwierigkeiten im Schriftspracherwerb. Möglichkeiten der Diagnose und Förderung*. Hamburg: Diplomica Verlag.

Gough, Philip B. & Tunmer, William E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education,* 7, S. 6-10.

Günther, Klaus B. (1995). Ein Stufenmodell der Entwicklung kindlicher Lese- und Schreibstrategien. In H. Balhorn & H. Brügelmann (Hrsg.), *Rätsel des Schriftspracherwebs. Neue Sichtweisen aus der Forschung* (S. 98-121). Lengwil am Bodensee: Libelle.

Günthner, Werner (2013). *Lesen und Schreiben lernen bei geistiger Behinderung. Grundlagen und Übungsvorschläge zum erweiterten Lese- und Schreibbegriff* (4. Auflage). Dortmund: Verlag modernes Lernen.

Hochstadt, Christiane, Krafft, Andreas & Olsen, Ralf (2015). *Deutschdidaktik. Konzeptionen für die Praxis* (2. Auflage). Tübingen: Narr Francke Attempto Verlag.

Holle, Karl (2009). Psychologische Lesemodelle und ihre lesedidaktischen Implikationen. In C. Garbe, K. Holle & T. Jesch (Hrsg.), *Texte lesen. Textverstehen Lesedidaktik Lesesozialisation* (2. Auflage, S. 103-166). Paderborn: Verlag Ferdinand Schöningh.

Holling, Heinz, Preckel, Franzis & Vock, Miriam (2004). *Intelligenzdiagnostik* (Band 6). Göttingen: Hogrefe Verlag.

Huber, Christian & Grosche, Michael (2012). Das response-to-intervention-Modell als Grundlage für einen inklusiven Paradigmenwechsel in der Sonderpädagogik. *Zeitschrift für Heilpädagogik*, 8, 312–322.

Jungjohann, Jana & Gebhardt, Markus (2018). Lernverlaufsdiagnostik im inklusiven Anfangsunterricht Lesen – Verschränkung von Lernverlaufsdiagnostik, Förderplanung und Wochenplanarbeit. In F. Hellmich, G. Görel, M. F. Löper (Hrsg.), *Inklusive Schul- und Unterrichtsentwicklung*, (S. 160-172). Stuttgart: Kohlhammer.

Jungjohann, Jana, Gebhardt, Markus, Diehl, Kirsten & Mühling, Andreas (2017). *Förderansätze im Lesen mit LEVUMI*. Online verfügbar unter: https://eldorado.tu-dortmund.de/bitstream/2003/36024/1/Förderansätze_Lehrerhandbuch%20LEVUMI.PDF [letzter Zugriff: 14.04.2018].

Jungjohann, Jana, Gegenfurtner, Andreas & Gebhardt, Markus (2018a). Systematisches Review von Lernverlaufsmessung im Bereich der frühen Leseflüssigkeit. *Empirische Sonderpädagogik*, 1, S. 100-118.

Jungjohann, Jana, DeVries, Jeffrey M., Gebhardt, Markus & Mühling, Andreas (2018b). Levumi: A Web-Based Curriculum-Based Measurement to Monitor Learning Progress in Inclusive Classrooms. In Miesenberger, K., Kouroupetroglou, G., Penaz, P. (Eds.), *Computers Helping People with Special Needs*. 16th International Conference, ICCHP 2018, Linz, Austria, July 2018, Proceedings. Wiesbaden: Springer.

Jungjohann, Jana, Rütter, Hanna & DeVries, Jeffrey (2018c). *How to SPSS für die Auswertung von Daten aus Levumi* (Version 1.0).

Kainz, Friedrich (1967). *Spezielle Sprachpsychologie* (Psychologie der Sprache, Band 4). Stuttgart: Ferdinand Enke Verlag.

Klauer, Karl Josef (2006). Erfassung des Lernfortschritts durch curriculumbasierte Messung. *Heilpädagogische Forschung*, 32, S. 16-26.

Klauer, Karl Josef (2011). Lernverlaufsdiagnostik – Konzept, Schwierigkeiten und Möglichkeiten. *Empirische Sonderpädagogik*, 3, S. 207-224.

Klauer, Karl Josef (2014). Formative Leistungsdiagnostik. Historischer Hintergrund und Weiterentwicklung zur Lernverlaufsdiagnostik. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (Tests und Trends, Band 12) (S. 1-18). Göttingen: Hogrefe Verlag.

Klicpera, Christian, Schabmann, Alfred & Gasteiger-Klicpera, Barbara (2013). *Legasthenie – LRS. Modelle, Diagnose, Therapie und Förderung*. München: Ernst Reinhardt Verlag.

Knigge, Michael (2015). Arten von Forschungsdesigns und Untersuchungsplänen. In K. Koch & S. Ellinger (Hrsg.), *Empirische Forschungsmethoden in der Heil- und Sonderpädagogik. Lehrbuch* (S. 57-67). Göttingen: Hogrefe Verlag.

Lienert, Gustav & Raatz, Ulrich (1998). *Testaufbau und Testanalyse* (6. Auflage). Weinheim: Psychologie Verlags Union.

Lord, Frederic M. & Novick, Melvin R. (1968). *Statistical Theories of Mental Test Scores*. Massachusetts: Addison-Wesley Publishing Company, Inc.

Maier, Uwe (2010). Formative Assessment – Ein erfolgversprechendes Konzept zur Reform von Unterricht und Leistungsmessung? *Zeitschrift für Erziehungswissenschaft*, 2, S. 293-308.

Maier, Uwe (2014). Computergestützte, formative Leistungsdiagnostik in Primar- und Sekundarschulen. Ein Forschungsüberblick zu Entwicklung, Implementation und Effekten. *Unterrichtswissenschaft*, 1, S. 69-86.

Maier, Uwe (2017). *Lehr-Lernprozesse in der Schule: Studium* (2. Auflage). Regensburg: Julius Klinkhardt.

Marx, Peter (2007). *Lese- und Rechtschreiberwerb*. Paderborn: Verlag Ferdinand Schöningh.

Ministerium für Schule und Weiterbildung - MSW (2008). *Richtlinien und Lehrpläne für die Grundschule in Nordrhein-Westfalen*. Frechen: Ritterbach Verlag.

Mühling, Andreas, Gebhardt, Markus & Diehl, Kirsten (2017). Formative Diagnostik durch die Onlineplattform Levumi. *Informatik Spectrum, 40 (6),* S. 556-561. Online verfügbar unter: https://link.springer.com/content/pdf/10.1007%2Fs00287-017-1069-7.pdf [letzter Zugriff: 14.04.2018].

Müller, B., Krizan, A., Hecht, T., Richter, T. & Ennemoser, M. (2013). Leseflüssigkeit im Grundschulalter: Entwicklungsverlauf und Effekte systematischer Leseförderung. *Lernen und Lernstörungen*, 2, S. 131-146.

Moosbrugger, Helfried & Kelava, Augustin (2008). *Testtheorie und Fragebogenkonstruktion*. Berlin u.a.: Springer.

Orthmann Bless, Dagmar (2015). Deskriptivstatistik und Inferenzstatistik. In K. Koch & S. Ellinger (Hrsg.), *Empirische Forschungsmethoden in der Heil- und Sonderpädagogik. Lehrbuch* (S. 106-112). Göttingen: Hogrefe Verlag.

Preckel, Franzis & Brüll, Matthias (2008). *Intelligenztests*. München: Ernst Reinhardt Verlag.

Pospeschill, Markus (2010). *Testtheorie, Testkonstruktion, Testevaluation.* München: Ernst Reinhardt Verlag.

Raithel, Jürgen (2008). *Quantitative Forschung. Ein Praxiskurs* (2. Auflage). Wiesbaden: VS Verlag für Sozialwissenschaften.

Richter, Karin & Plath, Monika (2012). *Lesemotivation in der Grundschule. Empirische Befunde und Modelle für den Unterricht* (3. Auflage). Weinheim und Basel: Beltz Juventa.

Rost, Jürgen (2004). *Lehrbuch. Testtheorie – Testkonstruktion* (2. Auflage). Bern: Verlag Hans Huber.

Scheerer-Neumann, Gerheid (2006a). Entwicklung der basalen Lesefähigkeit. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch.* Band I. (2. Auflage, S. 513-524). Paderborn: Ferdinand Schöningh.

Scheerer-Neumann, Gerheid (2006b). Leseschwierigkeiten. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch.* Band I. (2. Auflage, S. 551-567). Paderborn: Ferdinand Schöningh.

Scheerer-Neumann, Gerheid (2015). *Lese-Rechtschreib-Schwäche und Legasthenie. Grundlagen, Diagnostik und Förderung.* Stuttgart: W. Kohlhammer Verlag.

Schenk, Christa (2007). *Lesen und Schreiben lernen und lehren. Eine Didaktik des Schriftspracherwerbs* (7. Auflage). Baltmannsweiler: Schneider Verlag Hohengehren.

Schneider, Wolfgang (2017). *Lesen und Schreiben lernen. Wie erobern Kinder die Schriftsprache?* Würzburg: Springer Verlag.

Schründer-Lenzen, Agi (2009). *Schriftspracherwerb und Unterricht. Bausteine professionellen Handlungswissens* (3. Auflage). Wiesbaden: VS Verlag für Sozialwissenschaften.

Schwenk, Christin, Kuhn, Jörg-Tobias, Doebler, Philipp & Holling, Heinz (2017). Auf Goldmünzenjagd: Psychometrische Kennwerte verschiedener Scoringansätze bei computergestützter Lernverlaufsdiagnostik im Bereich Mathematik. *Empirische Sonderpädagogik*, 2, S. 123-142.

Scriven, Michael (1967). The Methodology of Evaluation. In R. W. Tyler, R. M. Gagné & M. Scriven (Eds.), *Perspectives of Curriculum Evaluation* (p. 39-83). Chicago: Rand McNally & Company.

Sikora, Simon (2015). Operationalisierung. In K. Koch & S. Ellinger (Hrsg.), *Empirische Forschungsmethoden in der Heil- und Sonderpädagogik. Lehrbuch* (S. 68- 75). Göttingen: Hogrefe Verlag.

Souvignier, Elmar & Förster, Natalie (2011). Effekte prozessorientierter Diagnostik auf die Entwicklung der Lesekompetenz leseschwacher Viertklässler. *Empirische Sonderpädagogik*, 3, S. 243-255.

Souvignier, Elmar, Förster, Natalie & Schulte, Elisabeth (2014a). Wirksamkeit formativen Assessments – Evaluation des Ansatzes der Lernverlaufsdiagnostik. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (Tests und Trends, Band 12) (S. 221-238). Göttingen: Hogrefe Verlag.

Souvignier, Elmar, Förster, Natalie & Salaschek, Martin (2014b). quop: Ein Ansatz internetbasierter Lernverlaufsdiagnostik mit Testkonzepten für Lesen und Mathematik. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (Tests und Trends, Band 12) (S. 239-256). Göttingen: Hogrefe Verlag.

Spinner, Kaspar H. (2010). Lesekompetenz ausbilden, Lesestandards erfüllen. G. Schulz (Hrsg.), *Lesen lernen in der Grundschule. Lesekompetenz und Leseverstehen. Förderung von Bücherwelten* (S. 48-61). Berlin: Cornelsen Verlag Scriptor.

Strathmann, Alfons M. & Klauer, Karl Josef (2010). Lernverlaufsdiagnostik: Ein Ansatz zur längerfristigen Lernfortschrittsmessung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 2, S. 111-122.

Strathmann, Alfons, Klauer, Karl Josef & Greisbach, Michaela (2010). Lernverlaufsdiagnostik – Dargestellt am Beispiel der Entwicklung der Rechtschreibkompetenz in der Grundschule. *Empirische Sonderpädagogik,* 1, S. 64-77.

Sturm, Afra (2011). Leseflüssigkeit als Bedingung fürs Textverstehen. *Alfa-Forum*, 76, S. 15-17.

Vock, Miriam, Preckel, Franzis & Holling, Heinz (2007). *Förderung Hochbegabter in der Schule. Evaluationsbefunde und Wirksamkeit von Maßnahmen.* Göttingen: Hogrefe Verlag.

Voß, Stefan (2013). *Curriculumbasierte Messverfahren im mathematischen Erstunterricht. Zur Güte und Anwendbarkeit einer Adaption US-amerikanischer Verfahren im deutschen Schulsystem*. Saarbrücken: Südwestdeutscher Verlag für Hochschulschriften. Online verfügbar unter: http://rosdok.uni-rostock.de/file/rosdok_disshab_0000001156/rosdok_derivate_0000005279/Dissertation_Voss_2014.pdf [letzter Zugriff: 22.06.2018).

Voß, Stefan & Hartke, Bodo (2014). Curriculumbasierte Messverfahren (CBM) als Methode der formativen Leistungsdiagnostik im RTI-Ansatz. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (Tests und Trends, Band 12) (S. 83-99). Göttingen: Hogrefe Verlag.

Voß, Stefan (2015). Mittelwerte, Modalwerte, Mediane. In K. Koch & S. Ellinger (Hrsg.), *Empirische Forschungsmethoden in der Heil- und Sonderpädagogik. Lehrbuch* (S. 123-128). Göttingen: Hogrefe Verlag.

Voß, Stefan & Gebhardt, Markus (2017). Schwerpunktthema: Verlaufsdiagnostik in der Schule. *Empirische Sonderpädagogik,* 2, S. 95-97

Walter, Jürgen (2008). Curriculumbasiertes Messen (CBM) als lernprozessbegleitende Diagnostik: Erste deutschsprachige Ergebnisse zur Validität, Reliabilität und Veränderungssensibilität eines robusten Indikators zur Lernfortschrittsmessung beim Lesen. *Heilpädagogische Forschung,* 2, S. 62-79.

Walter, Jürgen (2009). Theorie und Praxis Curriculumbasierten Messens (CBM) in Unterricht und Förderung. *Zeitschrift für Heilpädagogik*, 5, S. 162-171.

Walter, Jürgen (2010). Lernfortschrittsdiagnostik am Beispiel der Lesekompetenz (LDL): Messtechnische Grundlagen sowie Befunde über zu erwartende Zuwachsraten während der Grundschulzeit. *Heilpädagogische Forschung,* 4, S. 162-176.

Walter, Jürgen (2013). *VSL – Verlaufsdiagnostik sinnerfassendes Lesen*. Göttingen: Hogrefe Verlag.

Walter, Jürgen (2014). Lernfortschrittsdiagnostik Lesen (LDL) und Verlaufsdiagnostik sinnerfassenden Lesens (VSL): Zwei Verfahren als Instrumente einer formativ orientierten Lesediagnostik. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (Tests und Trends, Band 12) (S. 166-201). Göttingen: Hogrefe Verlag.

Wayman, Miya Miura, Wallace, Teri, Wiley, Hilda Ives, Tichá, Renáta & Espin, Christina A. (2007). Literature Synthesis on Curriculum-Based Measurement in Reading. *The Journal of Special Education,* 2, pp. 85-120.

Wember, Franz Bernhard (2012). Weiterführendes Lesen. In U. Heimlich & F. B. Wember (Hrsg.), *Didaktik des Unterrichts im Förderschwerpunkt Lernen. Ein Handbuch für Studium und Praxis* (2. Auflage, S. 191-205). Stuttgart: Kohlhammer.

Wilbert, Jürgen & Linnemann, Markus (2011). Kriterien zur Analyse eines Tests zur Lernverlaufsdiagnostik. *Empirische Sonderpädagogik*, 3, S. 225-242.

Wilbert, Jürgen (2014). Instrumente zur Lernverlaufsmessung: Gütekriterien und Auswertungsherausforderungen. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (Tests und Trends, Band 12) (S. 281-308). Göttingen: Hogrefe Verlag.

Wilckens, Susanne (2018). *Lese-Rechtschreib-Störung und Bildungsbiografie. Die Bedeutung des schulischen Schriftspracherwerbs für die Identitätsentwicklung*. Wiesbaden: Springer Fachmedien.

IV Tabellenverzeichnis

V Abbildungsverzeichnis

VI Anhang

Anhang A - Wörtliche Instruktionen zum sinnentnehmenden Lesetest

Anhang B - Grafiken zum sinnentnehmenden Lesetest

Anhang C - Grafiken zum Silbentest

Anhang D - Grafiken zum Wörtertest

Anhang E - Grafiken zum Pseudowörtertest

Anhang F - Tabellen zur Itemschwierigkeit – sinnentnehmendes Lesen

Anhang G - Tabellen zur Itemschwierigkeit – Wörter lesen

Anhang A

Wörtliche Instruktionen zum Sinnentnehmenden Lesetest

„Der kleine Drache LeVuMi möchte heute gerne wissen, wie gut du schon lesen kannst. LeVuMi hat viele Sätze mitgebracht, in denen immer ein Wort fehlt. LeVuMi fragt sich, ob du herausfinden kannst, welches Wort in den Satz passt?

Dafür arbeiten wir gleich zusammen am Computer. Am Computer siehst du nacheinander verschiedene Sätze. In jedem Satz fehlt immer genau ein Wort. Deine Aufgabe ist es, das richtige Wort herauszufinden. Dir werden dazu immer vier Wörter vorgeschlagen. Ein Wort passt und drei Wörter sind falsch. Mit der Maus kannst du die Wörter auswählen und weiterklicken. Du hast 7 Minuten lang Zeit, so viele Sätze zu lesen, wie du schaffst. Konzentriere dich gut, damit du keine Fehler machst. In drei Wochen komme ich noch einmal wieder und du darfst noch einmal mit LeVuMi lesen. Dann zeigt LeVuMi dir, ob du etwas besser lesen kannst als heute.“

Ortswechsel zum Computer. Entweder den Log-In gemeinsam mit den Kindern durchführen. Alternativ können die Schüleraccounts auch schon vorher geöffnet werden. Bitte überprüfen Sie, dass jedes Kind am richtigen Account sitzt und arbeitet. Wenn alle Kinder die Übersichtsseite sehen, kann mit der Erklärung fortgefahren werden.

„Hier siehst du ein Beispiel. Der Satz lautet: Ein … kann fahren. Dazu siehst du vier Wörter zur Auswahl. Hier sind es: Auge, Essen, Sonntag, Auto. Welches Wort passt in den Satz?“

Eine Schülerin oder einen Schüler antworten lassen.

„Genau. Ein Auto kann fahren. Wenn du weißt, welches Wort richtig ist, kannst du es mit der Maus auswählen. Dazu musst du es nur anklicken. Danach siehst du das Wort in der Lücke. Wenn du dich um entscheidest, kannst du danach auch ein anderes Wort auswählen. Dafür klickst du das andere Wort wieder mit der Maus an. Dir werden immer vier Wörter vorgeschlagen. Ein Wort passt in den Satz. Die anderen drei sind falsch und ergeben keinen Sinn. Wenn du dir sicher bist, dass du das richtige Wort ausgewählt hast und es in der Lücke steht, klickst du auf ‚weiter‘. Danach zeigt dir LeVuMi einen neuen Satz an, in dem wieder ein Wort fehlt. Denke daran, dass du 7 Minuten Zeit hast. Der Computer achtet für dich auf die Zeit. Die Zeit ist um, wenn du LeVuMi in groß siehst. Hast du noch eine Frage?“

Fragen der Schülerinnen und Schüler klären.

„Dann kannst du jetzt auf Start klicken. Danach zeigt LeVuMi dir direkt den ersten Satz.“

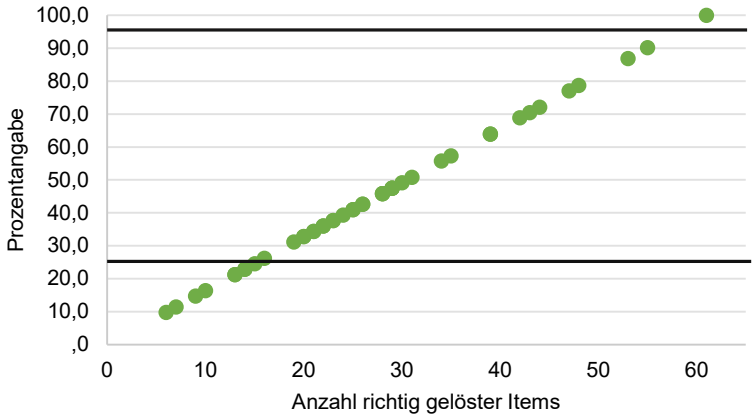
Anhang B

Grafiken zum sinnentnehmenden Lesetest

Jahrgangsstufe 3

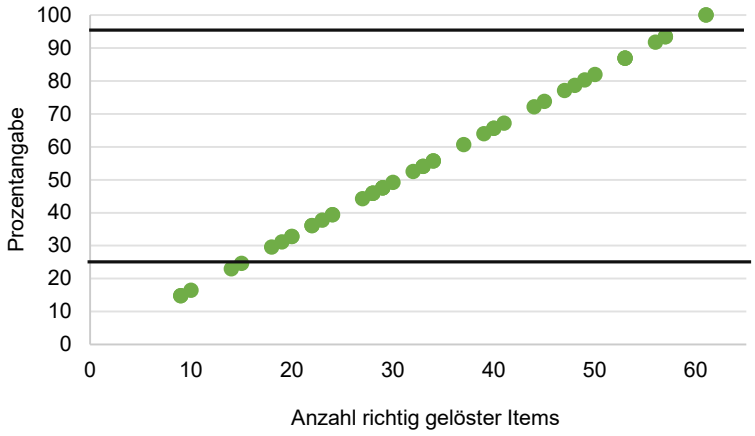
Messzeitpunkt 1

Werte - MZP 1	
Stichprobengröße	44
Spannweite	55
Minimalwert	6
Erstes Quartil	19,75
Median	25,5
Drittes Quartil	34,25
Maximalwert	61
Mittelwert	27,41
Modalwert	29
Standardabweichung	13,01



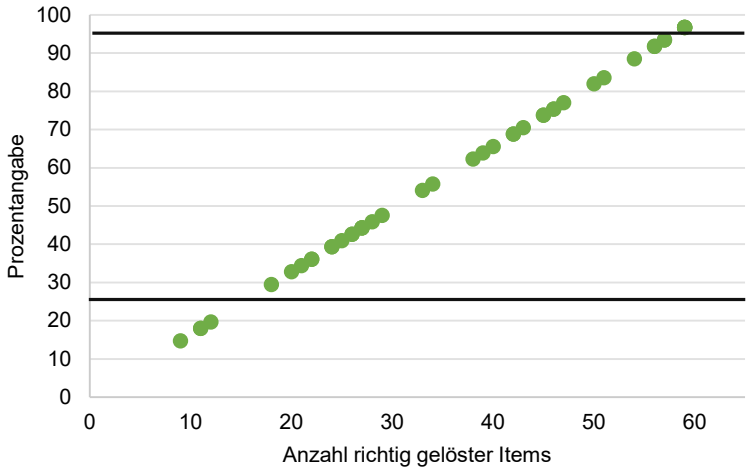
Messzeitpunkt 2

Werte – MZP2	
Stichprobengröße	46
Spannweite	52
Minimalwert	9
Erstes Quartil	23,25
Median	32,5
Drittes Quartil	46,5
Maximalwert	61
Mittelwert	34,22
Modalwert	29
Standardabweichung	14,47



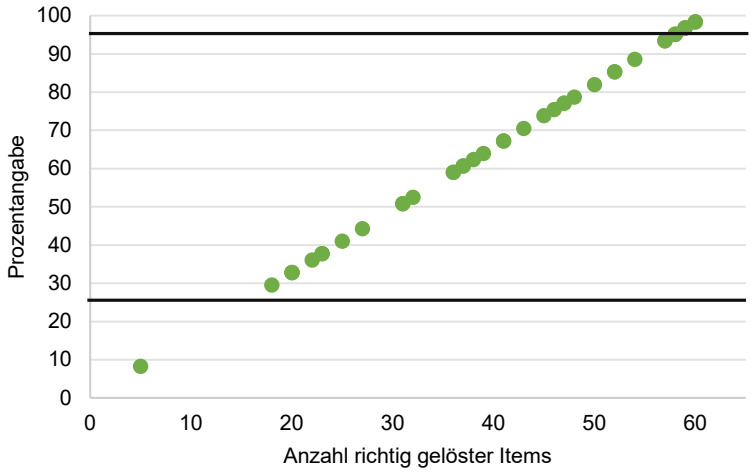
Messzeitpunkt 3

Werte – MZP3	
Stichprobengröße	44
Spannweite	50
Minimalwert	9
Erstes Quartil	24,75
Median	38,5
Drittes Quartil	50,25
Maximalwert	59
Mittelwert	36,93
Modalwert	59
Standardabweichung	15,37



Messzeitpunkt 4

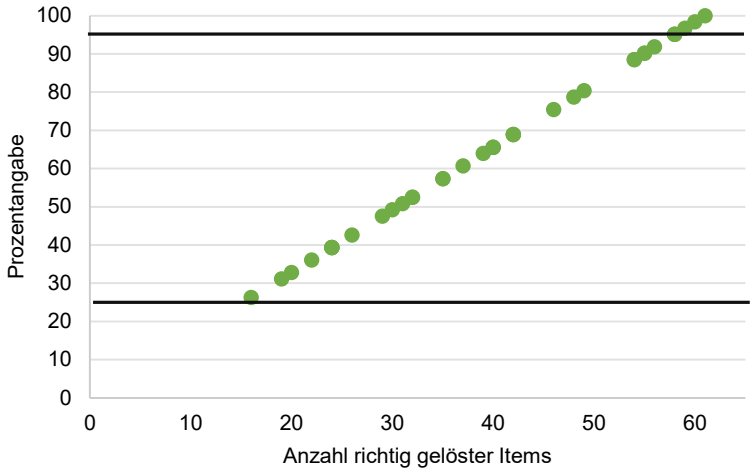
Werte – MZP4	
Stichprobengröße	42
Spannweite	55
Minimalwert	5
Erstes Quartil	31
Median	42
Drittes Quartil	56,25
Maximalwert	60
Mittelwert	41,17
Modalwert	58
Standardabweichung	14,64



Jahrgangsstufe 4

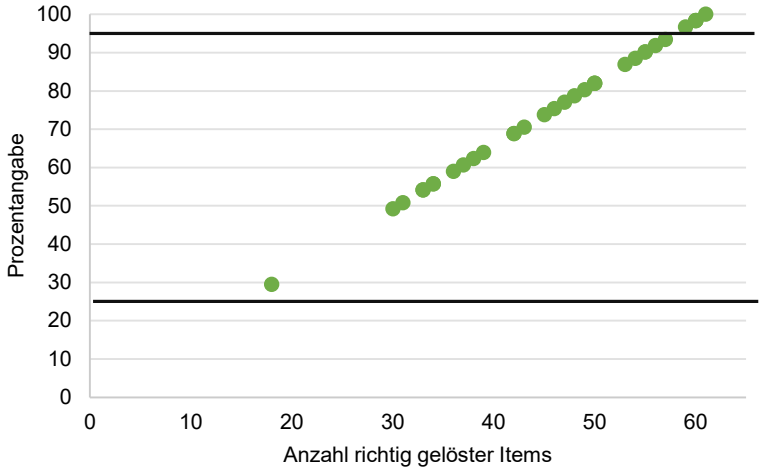
Messzeitpunkt 1

Werte - MZP1	
Stichprobengröße	38
Spannweite	45
Minimalwert	16
Erstes Quartil	29,25
Median	40
Drittes Quartil	54
Maximalwert	61
Mittelwert	39,89
Modalwert	24
Standardabweichung	13,5



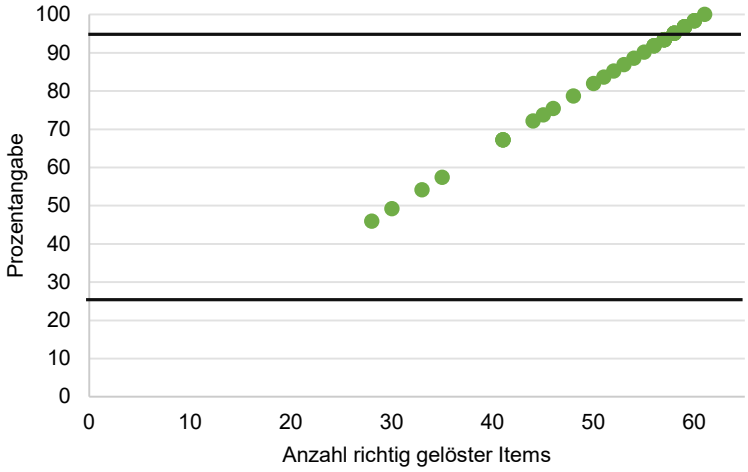
Messzeitpunkt 2

Werte – MZP2	
Stichprobengröße	33
Spannweite	43
Minimalwert	18
Erstes Quartil	37
Median	47
Drittes Quartil	55
Maximalwert	61
Mittelwert	45,61
Modalwert	60
Standardabweichung	10,68



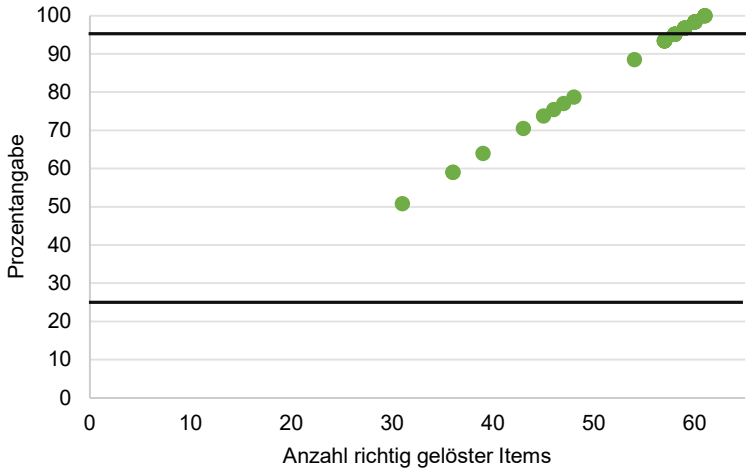
Messzeitpunkt 3

Werte – MZP3	
Stichprobengröße	40
Spannweite	33
Minimalwert	28
Erstes Quartil	47,5
Median	56,5
Drittes Quartil	58,25
Maximalwert	61
Mittelwert	52,1
Modalwert	59
Standardabweichung	8,9



Messzeitpunkt 4

Werte – MZP4	
Stichprobengröße	40
Spannweite	30
Minimalwert	31
Erstes Quartil	56,25
Median	58
Drittes Quartil	59
Maximalwert	61
Mittelwert	54,63
Modalwert	57
Standardabweichung	7,82



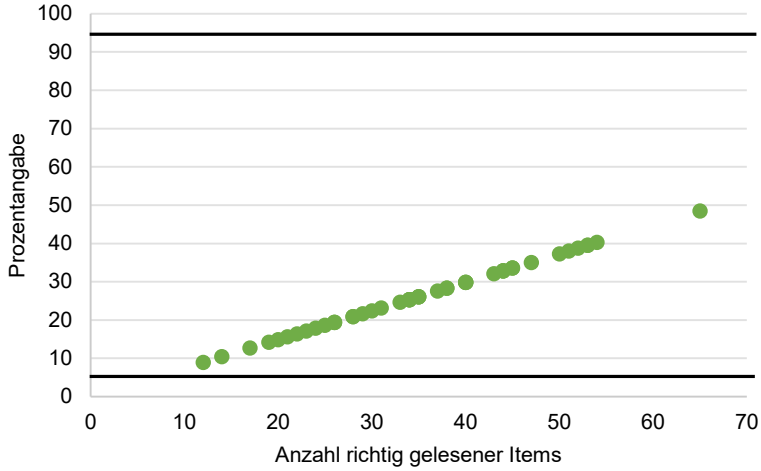
Anhang C

Grafiken zum Silbentest

Jahrgangsstufe 3

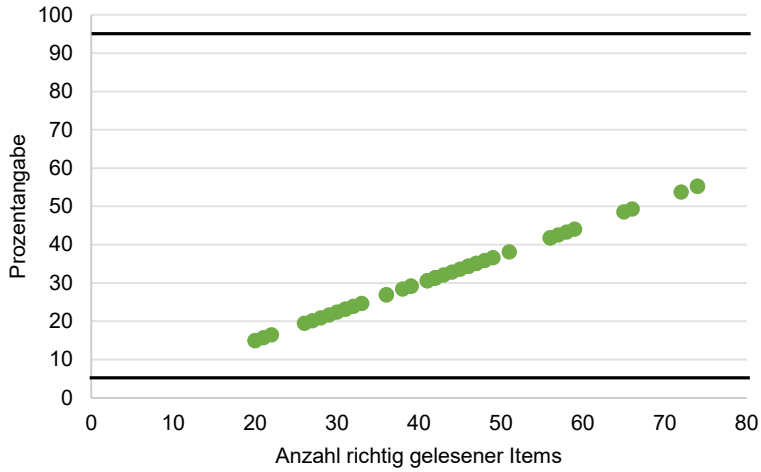
Messzeitpunkt 1

Werte – MZP 1	
Stichprobengröße	46
Spannweite	53
Minimalwert	12
Erstes Quartil	26
Median	35
Drittes Quartil	44
Maximalwert	65
Mittelwert	35
Modalwert	34
Standardabweichung	11,76



Messzeitpunkt 2

Werte – MZP 2	
Stichprobengröße	45
Spannweite	54
Minimalwert	20
Erstes Quartil	32
Median	42
Drittes Quartil	49
Maximalwert	74
Mittelwert	43
Modalwert	42
Standardabweichung	13,96

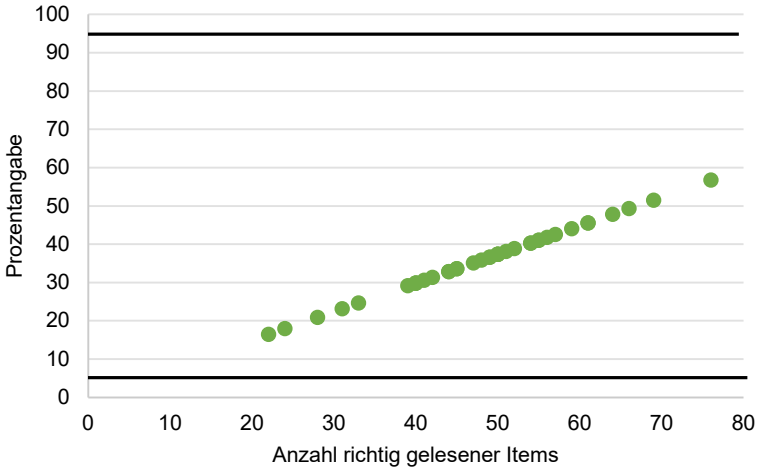


Messzeitpunkt 3

Werte – MZP 3	
Stichprobengröße	46
Spannweite	50
Minimalwert	22
Erstes Quartil	37
Median	46
Drittes Quartil	55
Maximalwert	72
Mittelwert	46
Modalwert	37
Standardabweichung	13,28

Messzeitpunkt 4

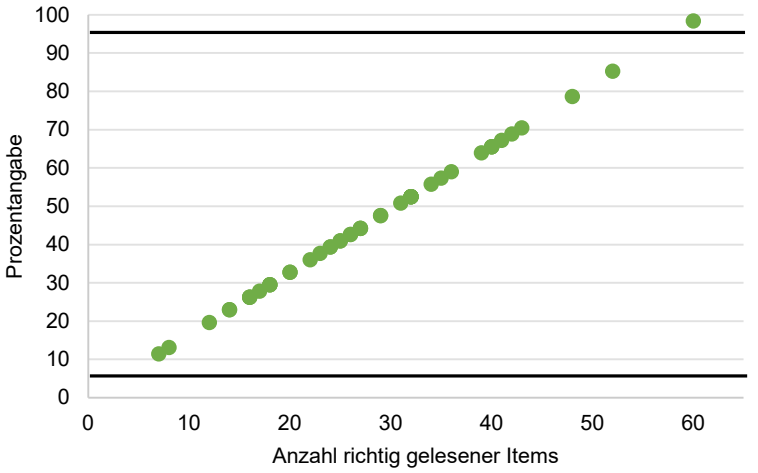
Werte – MZP 4	
Stichprobengröße	48
Spannweite	54
Minimalwert	22
Erstes Quartil	41,25
Median	50
Drittes Quartil	55
Maximalwert	76
Mittelwert	48
Modalwert	50
Standardabweichung	11,69



Jahrgangsstufe 4

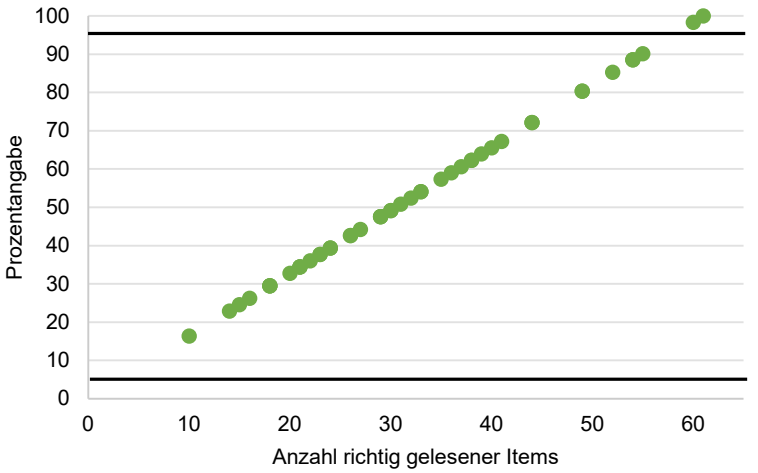
Messzeitpunkt 1

Werte – MZP 1	
Stichprobengröße	24
Spannweite	64
Minimalwert	17
Erstes Quartil	36,75
Median	44
Drittes Quartil	52
Maximalwert	81
Mittelwert	44,25
Modalwert	52
Standardabweichung	12,94



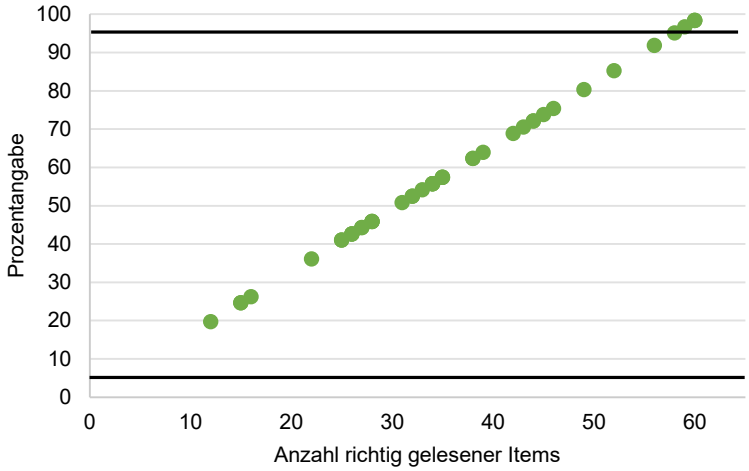
Messzeitpunkt 2

Werte – MZP 2	
Stichprobengröße	38
Spannweite	55
Minimalwert	18
Erstes Quartil	41,5
Median	48,5
Drittes Quartil	55
Maximalwert	73
Mittelwert	48,71
Modalwert	55
Standardabweichung	11,43



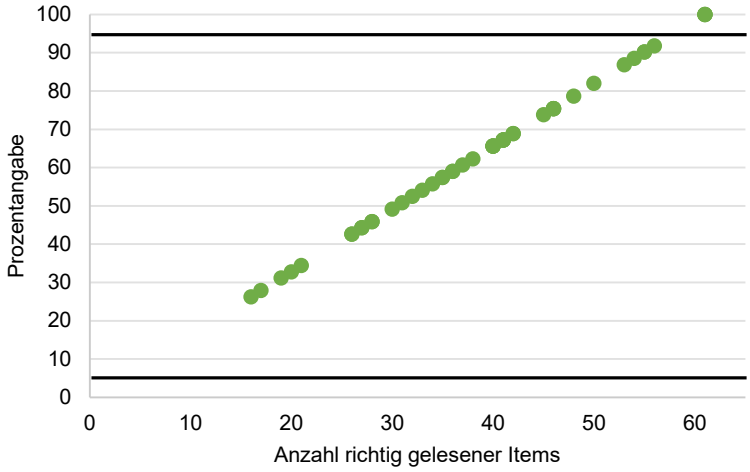
Messzeitpunkt 3

Werte – MZP 3	
Stichprobengröße	40
Spannweite	49
Minimalwert	35
Erstes Quartil	49
Median	55
Drittes Quartil	63
Maximalwert	84
Mittelwert	55,85
Modalwert	55
Standardabweichung	10,4



Messzeitpunkt 4

Werte – MZP 4	
Stichprobengröße	35
Spannweite	45
Minimalwert	38
Erstes Quartil	54,5
Median	61
Drittes Quartil	66,5
Maximalwert	83
Mittelwert	60,03
Modalwert	61
Standardabweichung	10,12



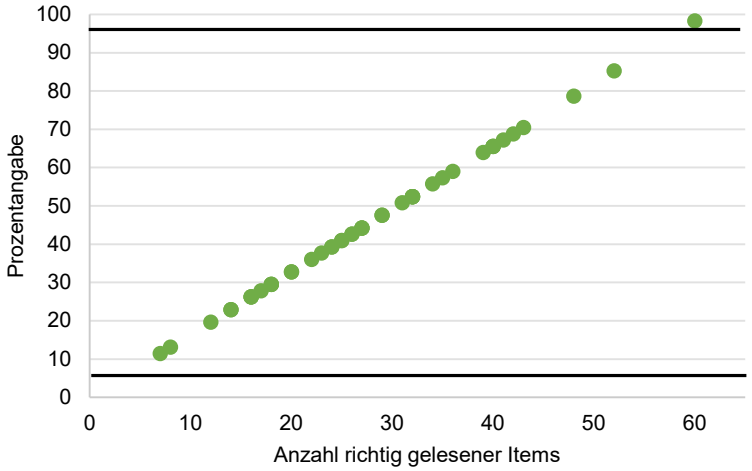
Anhang D

Grafiken zum Wörtertest

Jahrgangsstufe 3

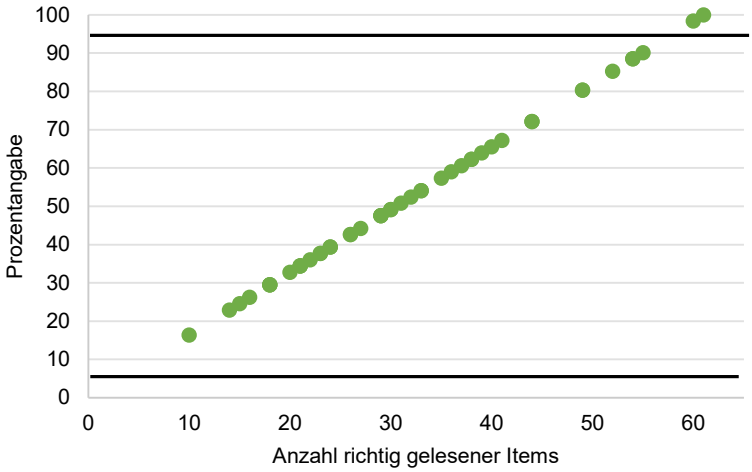
Messzeitpunkt 1

Werte – MZP 1	
Stichprobengröße	46
Spannweite	53
Minimalwert	7
Erstes Quartil	18,5
Median	27
Drittes Quartil	34,75
Maximalwert	60
Mittelwert	28,02
Modalwert	32
Standardabweichung	11,4



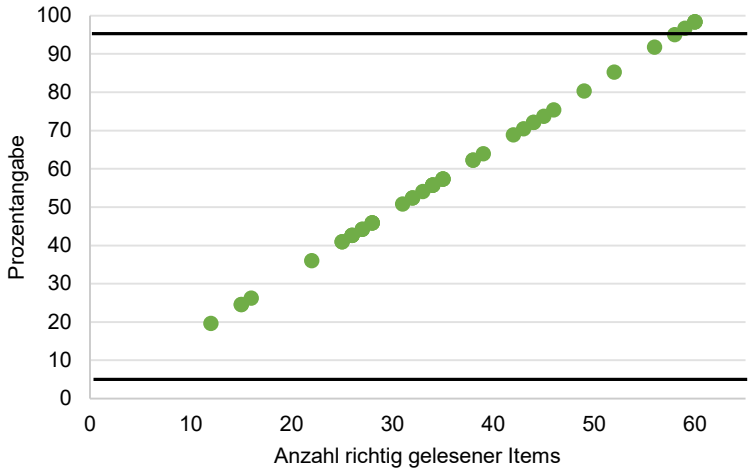
Messzeitpunkt 2

Werte – MZP 2	
Stichprobengröße	46
Spannweite	51
Minimalwert	10
Erstes Quartil	22,25
Median	30
Drittes Quartil	39,75
Maximalwert	61
Mittelwert	32,37
Modalwert	21
Standardabweichung	12,91



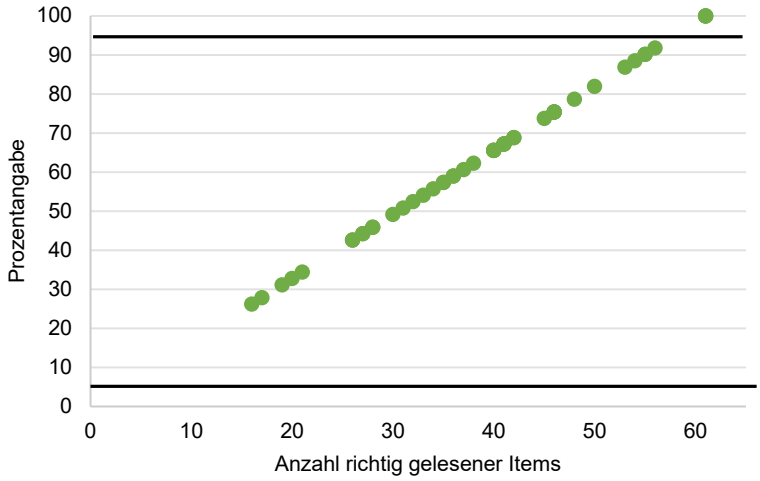
Messzeitpunkt 3

Werte – MZP 3	
Stichprobengröße	43
Spannweite	48
Minimalwert	12
Erstes Quartil	27
Median	34
Drittes Quartil	44
Maximalwert	60
Mittelwert	36
Modalwert	34
Standardabweichung	12,77



Messzeitpunkt 4

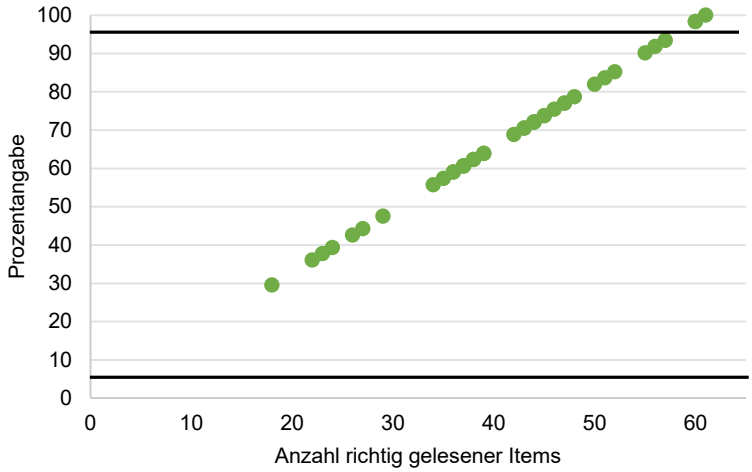
Werte – MZP 4	
Stichprobengröße	47
Spannweite	45
Minimalwert	16
Erstes Quartil	30,5
Median	40
Drittes Quartil	46
Maximalwert	61
Mittelwert	38,79
Modalwert	41
Standardabweichung	11,69



Jahrgangsstufe 4

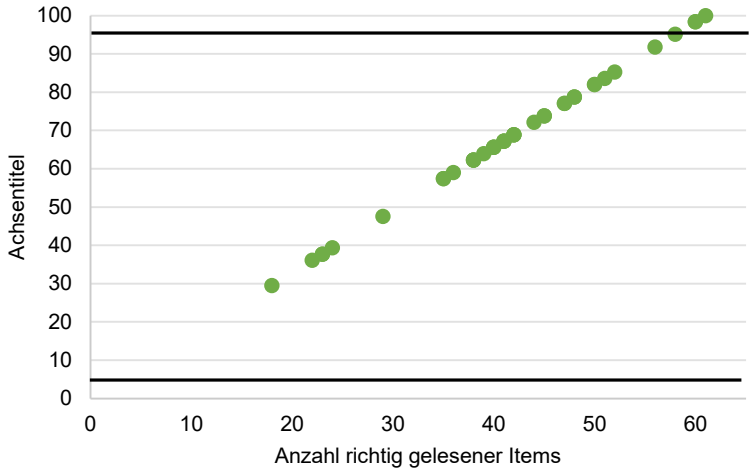
Messzeitpunkt 1

Werte – MZP 1	
Stichprobengröße	39
Spannweite	43
Minimalwert	18
Erstes Quartil	36
Median	43
Drittes Quartil	47,5
Maximalwert	61
Mittelwert	41,41
Modalwert	44
Standardabweichung	10,84



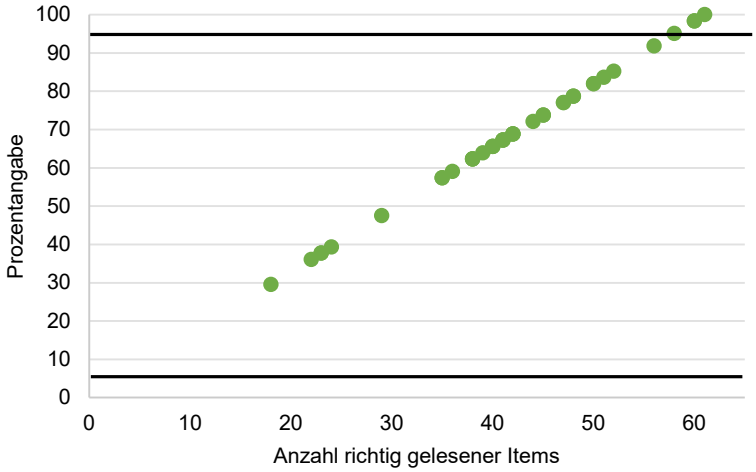
Messzeitpunkt 2

Werte – MZP 2	
Stichprobengröße	38
Spannweite	43
Minimalwert	18
Erstes Quartil	38
Median	41,5
Drittes Quartil	48
Maximalwert	61
Mittelwert	41,82
Modalwert	42
Standardabweichung	10,65



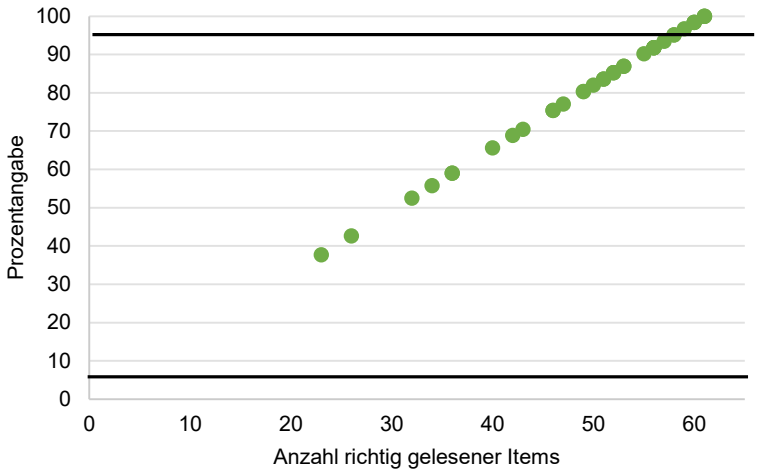
Messzeitpunkt 3

Werte – MZP 3	
Stichprobengröße	40
Spannweite	37
Minimalwert	24
Erstes Quartil	41,75
Median	48,5
Drittes Quartil	54
Maximalwert	61
Mittelwert	46,93
Modalwert	43
Standardabweichung	9,16



Messzeitpunkt 4

Werte – MZP 4	
Stichprobengröße	37
Spannweite	38
Minimalwert	23
Erstes Quartil	46
Median	52
Drittes Quartil	57
Maximalwert	61
Mittelwert	49,65
Modalwert	56
Standardabweichung	9,93



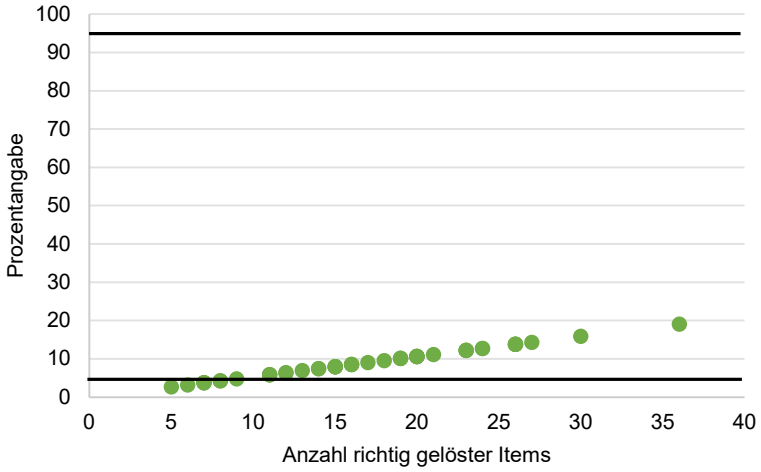
Anhang E

Grafiken zum Pseudowörtertest

Jahrgangsstufe 3

Messzeitpunkt 1

Werte – MZP 1	
Stichprobengröße	46
Spannweite	31
Minimalwert	5
Erstes Quartil	12
Median	16,5
Drittes Quartil	22,5
Maximalwert	36
Mittelwert	17,17
Modalwert	20
Standardabweichung	6,97

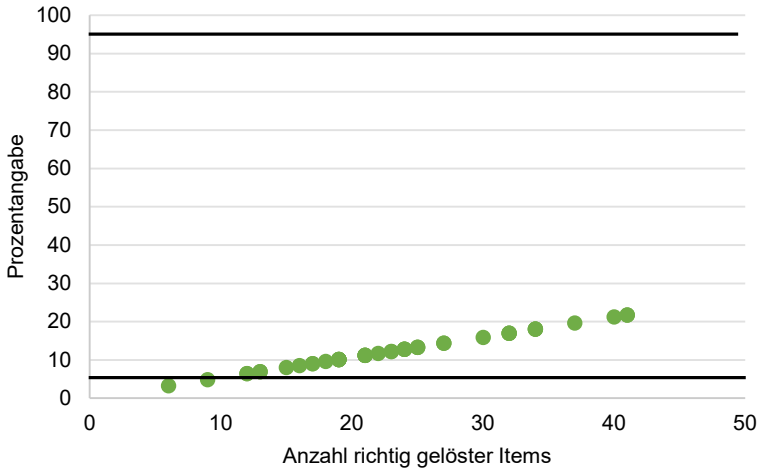


Messzeitpunkt 2

Werte – MZP 2	
Stichprobengröße	45
Spannweite	36
Minimalwert	7
Erstes Quartil	13
Median	20
Drittes Quartil	25
Maximalwert	43
Mittelwert	20,11
Modalwert	13
Standardabweichung	8,32

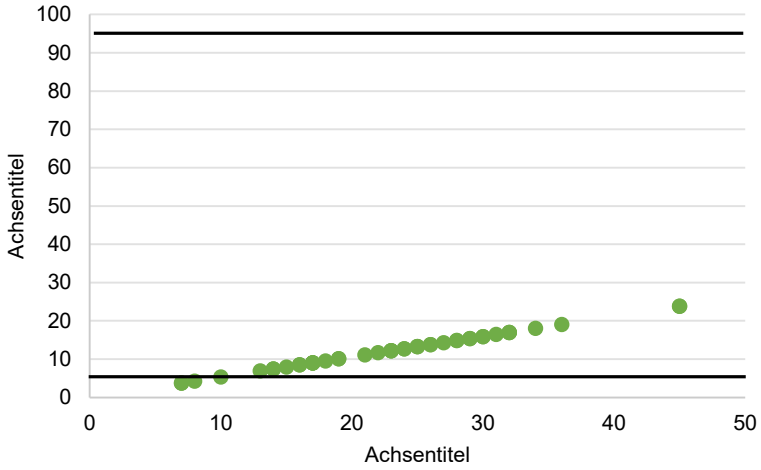
Messzeitpunkt 3

Werte	
Stichprobengröße	44
Spannweite	35
Minimalwert	6
Erstes Quartil	18,75
Median	22,5
Drittes Quartil	27,75
Maximalwert	41
Mittelwert	23,32
Modalwert	19
Standardabweichung	8,37



Messzeitpunkt 4

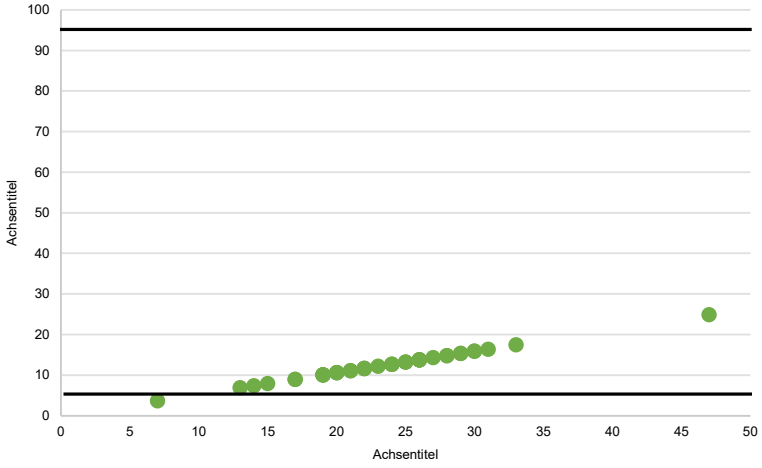
Werte	
Stichprobengröße	44
Spannweite	38
Minimalwert	7
Erstes Quartil	17
Median	23,5
Drittes Quartil	29
Maximalwert	45
Mittelwert	23,16
Modalwert	17
Standardabweichung	8,75



Jahrgangsstufe 4

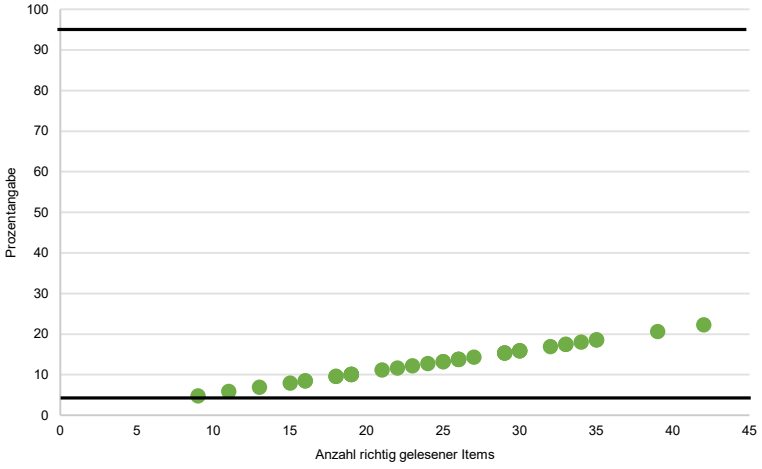
Messzeitpunkt 1

Werte – MZP 1	
Stichprobengröße	39
Spannweite	40
Minimalwert	7
Erstes Quartil	19,5
Median	23
Drittes Quartil	26,5
Maximalwert	47
Mittelwert	23,33
Modalwert	26
Standardabweichung	6,59



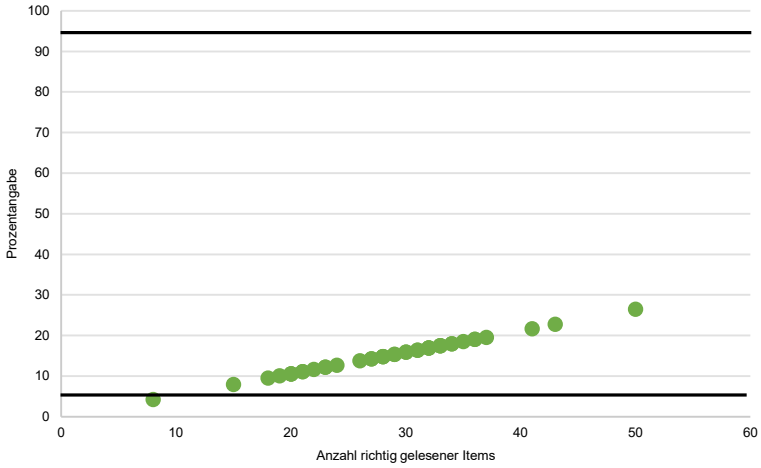
Messzeitpunkt 2

Werte – MZP 2	
Stichprobengröße	38
Spannweite	33
Minimalwert	9
Erstes Quartil	19
Median	25,5
Drittes Quartil	30
Maximalwert	42
Mittelwert	24,79
Modalwert	19
Standardabweichung	7,68



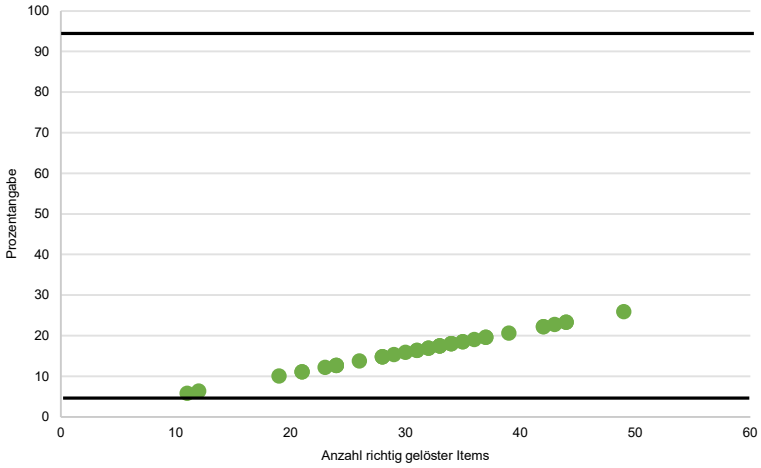
Messzeitpunkt 3

Werte – MZP 3	
Stichprobengröße	40
Spannweite	42
Minimalwert	8
Erstes Quartil	22,75
Median	29
Drittes Quartil	33
Maximalwert	50
Mittelwert	28,48
Modalwert	33
Standardabweichung	7,9



Messzeitpunkt 4

Werte – MZP 4	
Stichprobengröße	37
Spannweite	38
Minimalwert	11
Erstes Quartil	26
Median	32
Drittes Quartil	36
Maximalwert	49
Mittelwert	31,3
Modalwert	28
Standardabweichung	8,47



Anhang F

Tabellen zur Itemschwierigkeit – sinnentnehmendes Lesen

Jahrgangsstufe 3

	MZP 1		MZP 2		MZP 3[7]		MZP 4	
Items	Mittelwert	P_i	Mittelwert	P_i	Mittelwert	P_i	Mittelwert	P_i
Augen	0,89	88,64	0,52	52,17	0,57	56,82	0,76	76,19
fleißig	0,93	93,18	0,65	65,22	0,52	52,27	0,71	71,43
Sobald	0,77	77,27	0,50	50,00	0,68	68,18	0,69	69,05
Blumen	0,89	88,64	0,59	58,70	0,59	59,09	0,83	83,33
schmeckt	0,89	88,64	0,63	63,04	0,50	50,00	0,71	71,43
In	0,36	36,36	0,17	17,39	0,16	15,91	0,38	38,10
Schere	0,86	86,36	0,63	63,04	0,66	65,91	0,81	80,95
Haus	0,91	90,91	0,48	47,83	0,50	50,00	0,71	71,43
wegen	0,77	77,27	0,54	54,35	0,70	70,45	0,69	69,05
bevor	0,75	75,00	0,46	45,65	0,75	75,00	0,64	64,29
wohne	0,80	79,55	0,72	71,74	0,55	54,55	0,71	71,43
schnelle	0,82	81,82	0,67	67,39	0,55	54,55	0,67	66,67
aber	0,77	77,27	0,48	47,83	0,77	77,27	0,69	69,05
böse	0,86	86,36	0,65	65,22	0,66	65,91	0,69	69,05
Baby	0,84	84,09	0,57	56,52	0,66	65,91	0,67	66,67
Zwischen	0,68	68,18	0,59	58,70	0,61	61,36	0,64	64,29
spricht	0,73	72,73	0,52	52,17	0,36	36,36	0,67	66,67
Bett	0,82	81,82	0,59	58,70	0,73	72,73	0,76	76,19
von	0,80	79,55	0,50	50,00	0,61	61,36	0,71	71,43
scheint	0,59	59,09	0,61	60,87	0,50	50,00	0,71	71,43
Nachdem	0,59	59,09	0,43	43,48	0,64	63,64	0,64	64,29
runde	0,80	79,55	0,61	60,87	0,68	68,18	0,74	73,81
Freunde	0,70	70,45	0,61	60,87	0,64	63,64	0,69	69,05
gut	0,73	72,73	0,72	71,74	0,61	61,36	0,79	78,57
Vogel	0,61	61,36	0,54	54,35	0,68	68,18	0,67	66,67
Anstatt	0,48	47,73	0,46	45,65	0,55	54,55	0,67	66,67
lustigen	0,57	56,82	0,63	63,04	0,57	56,82	0,71	71,43
Durch	0,45	45,45	0,39	39,13	0,68	68,18	0,62	61,90
Beine	0,55	54,55	0,59	58,70	0,61	61,36	0,79	78,57
sammeln	0,48	47,73	0,72	71,74	0,59	59,09	0,67	66,67
hungrig	0,52	52,27	0,67	67,39	0,57	56,82	0,62	61,90
weder	0,32	31,82	0,43	43,48	0,50	50,00	0,57	57,14
Sonne	0,45	45,45	0,57	56,52	0,64	63,64	0,67	66,67
unter	0,39	38,64	0,61	60,87	0,82	81,82	0,62	61,90
schläft	0,43	43,18	0,61	60,87	0,57	56,82	0,62	61,90
Schuhe	0,34	34,09	0,48	47,83	0,55	54,55	0,64	64,29
backt	0,34	34,09	0,61	60,87	0,59	59,09	0,74	73,81
über	0,18	18,18	0,35	34,78	0,41	40,91	0,40	40,48
Bilder	0,27	27,27	0,54	54,35	0,59	59,09	0,67	66,67
Wenn	0,16	15,91	0,37	36,96	0,73	72,73	0,48	47,62
süß	0,25	25,00	0,67	67,39	0,59	59,09	0,74	73,81
Frösche	0,18	18,18	0,59	58,70	0,66	65,91	0,74	73,81
Schwester	0,20	20,45	0,63	63,04	0,66	65,91	0,74	73,81
dicke	0,20	20,45	0,61	60,87	0,48	47,73	0,64	64,29
Biene	0,14	13,64	0,46	45,65	0,55	54,55	0,71	71,43
Während	0,14	13,64	0,50	50,00	0,73	72,73	0,62	61,90
für	0,09	9,09	0,46	45,65	0,70	70,45	0,74	73,81
isst	0,16	15,91	0,65	65,22	0,64	63,64	0,71	71,43
Enten	0,14	13,64	0,63	63,04	0,64	63,64	0,76	76,19
Auf	0,14	13,64	0,52	52,17	0,73	72,73	0,69	69,05
wartest	0,11	11,36	0,52	52,17	0,45	45,45	0,64	64,29
Wasser	0,07	6,82	0,61	60,87	0,59	59,09	0,76	76,19
neues	0,07	6,82	0,67	67,39	0,64	63,64	0,67	66,67

[7] Wie bereits in der Ergebnisdarstellung darauf hingewiesen, fallen zu Messzeitpunkt 3 der Jahrgangsstufe 3 einige Präpositionen und Junktoren auffallend leichter aus als zu den anderen Messzeitpunkten. Diese sind in der Tabelle durch die grüne Farbe hervorgehoben.

außer	0,07	6,82	0,46	45,65	0,66	65,91	0,67	66,67
spitz	0,07	6,82	0,72	71,74	0,59	59,09	0,62	61,90
weil	0,05	4,55	0,50	50,00	0,59	59,09	0,60	59,52
kauft	0,07	6,82	0,70	69,57	0,64	63,64	0,67	66,67
Hase	0,05	4,55	0,70	69,57	0,61	61,36	0,69	69,05
Tür	0,05	4,55	0,54	54,35	0,66	65,91	0,67	66,67
mit	0,05	4,55	0,41	41,30	0,55	54,55	0,55	54,76
Büro	0,05	4,55	0,67	67,39	0,75	75,00	0,67	66,67
Mittelwert	0,45	44,93	0,56	56,09	0,61	60,54	0,67	67,49

Jahrgangsstufe 4

	MZP 1		MZP 2		MZP 3		MZP 4	
Items	Mittelwert	P_i	Mittelwert	P_i	Mittelwert	P_i	Mittelwert	P_i
Augen	0,68	68,42	0,76	75,76	0,85	85	0,875	87,5
fleißig	0,71	71,05	0,7	69,7	0,875	87,5	0,9	90
Sobald	0,58	57,89	0,79	78,79	0,825	82,5	0,9	90
Blumen	0,66	65,79	0,67	66,67	0,925	92,5	0,875	87,5
schmeckt	0,66	65,79	0,76	75,76	0,85	85	0,9	90
In	0,34	34,21	0,33	33,33	0,4	40	0,625	62,5
Schere	0,76	76,32	0,79	78,79	0,9	90	0,925	92,5
Haus	0,68	68,42	0,72	72,73	0,85	85	0,9	90
wegen	0,74	73,68	0,79	78,79	0,9	90	0,8	80
bevor	0,61	60,53	0,82	81,82	0,9	90	0,85	85
wohne	0,66	65,79	0,88	87,88	0,975	97,5	0,925	92,5
schnelle	0,71	71,05	0,79	78,79	0,85	85	0,9	90
aber	0,61	60,53	0,85	84,85	0,9	90	0,95	95
böse	0,66	65,79	0,82	81,82	0,85	85	0,975	97,5
Baby	0,63	63,16	0,7	69,7	0,9	90	0,975	97,5
Zwischen	0,61	60,53	0,67	66,67	0,85	85	0,85	85
spricht	0,61	60,53	0,73	72,73	0,825	82,5	0,8	80
Bett	0,82	81,58	0,73	72,73	0,925	92,5	0,925	92,5
von	0,58	57,89	0,76	75,76	0,85	85	0,9	90
scheint	0,63	63,16	0,7	69,7	0,825	82,5	0,875	87,5
Nachdem	0,55	55,26	0,7	69,7	0,8	80	0,875	87,5
runde	0,63	63,16	0,73	72,73	1	100	0,9	90
Freunde	0,68	68,42	0,73	72,723	0,9	90	0,85	85
gut	0,74	73,68	0,85	84,85	0,85	85	0,925	92,5
Vogel	0,76	76,32	0,67	66,67	0,875	87,5	0,95	95
Anstatt	0,61	60,53	0,64	63,64	0,725	72,5	0,825	82,5
lustigen	0,71	71,05	0,79	78,79	0,925	92,5	0,95	95
Durch	0,66	65,79	0,67	66,67	0,725	72,5	0,8	80
Beine	0,66	65,79	0,73	72,73	0,975	97,5	0,9	90
sammeln	0,68	68,42	0,88	87,88	0,875	87,5	0,95	95
hungrig	0,66	65,79	0,73	72,73	0,9	90	1	100
weder	0,55	55,26	0,73	72,73	0,725	72,5	0,875	87,5
Sonne	0,63	63,16	0,79	78,79	0,85	85	0,95	95
unter	0,76	76,32	0,82	81,82	0,85	85	0,9	90
schläft	0,66	65,79	0,79	78,79	0,8	80	0,975	97,5
Schuhe	0,71	71,05	0,76	75,76	0,9	90	0,9	90
backt	0,66	65,79	0,76	75,76	0,825	82,5	0,9	90
über	0,34	34,21	0,61	60,61	0,7	70	0,775	77,5
Bilder	0,68	68,42	0,7	69,7	0,9	90	0,875	87,5
Wenn	0,63	63,16	0,76	75,76	0,75	75	0,875	87,5
süß	0,68	68,42	0,82	81,82	0,925	92,5	0,95	95
Frösche	0,68	68,42	0,73	72,73	0,875	87,5	0,9	90
Schwester	0,68	68,42	0,79	78,79	0,85	85	0,9	90
dicke	0,68	68,42	0,67	66,67	0,825	82,5	0,775	77,5
Biene	0,68	68,42	0,64	63,64	0,925	92,5	0,875	87,5
Während	0,63	63,16	0,82	81,82	0,8	80	0,85	85
für	0,66	65,79	0,73	72,73	0,8	80	0,875	87,5
isst	0,63	63,16	0,88	87,88	0,85	85	0,925	92,5
Enten	0,76	76,32	0,76	75,76	0,925	92,5	0,925	92,5
Auf	0,66	65,79	0,85	84,85	0,875	87,5	0,925	92,5
wartest	0,66	65,79	0,76	75,76	0,85	85	0,925	92,5
Wasser	0,68	68,42	0,82	81,82	0,9	90	0,95	95
neues	0,71	71,05	0,79	78,79	0,925	92,5	0,95	95
außer	0,66	65,79	0,79	78,79	0,775	77,5	0,875	87,5
spitz	0,76	76,32	0,82	81,82	0,875	87,5	0,925	92,5
weil	0,53	52,63	0,67	69,7	0,825	82,5	0,925	92,5
kauft	0,61	60,53	0,73	72,73	0,9	90	0,95	95
Hase	0,76	76,32	0,79	78,79	0,85	85	0,925	92,5
Tür	0,66	65,79	0,73	72,73	0,975	97,5	0,95	95
mit	0,58	57,89	0,73	72,73	0,825	82,5	0,85	85
Büro	0,63	63,16	0,79	78,79	0,875	87,5	0,925	92,5
Mittelwert	0,65	65,4	0,74	74,76	0,85	85,41	0,9	89,55

Anhang G

Tabellen zur Itemschwierigkeit – Wörter lesen

Jahrgangsstufe 3

	MZP 1		MZP 2		MZP 3		MZP 4	
Items	Mittelwert	P_i	Mittelwert	P_i	Mittelwert	P_i	Mittelwert	P_i
Gleis	0,96	95,65	0,59	58,7	0,52	52,27	0,87	87,23
schreiben	1	100	0,59	58,7	0,61	61,36	0,85	85,11
Wolke	0,98	97,83	0,43	43,48	0,68	68,18	0,89	89,36
Blüte	0,93	93,48	0,52	52,17	0,48	47,73	0,79	78,72
Rinder	0,89	89,13	0,43	43,48	0,73	72,73	0,87	87,23
Forst	0,63	63,04	0,52	52,17	0,41	40,91	0,72	72,34
Traube	0,93	93,48	0,52	52,17	0,61	61,36	0,87	87,23
Durst	0,96	95,65	0,63	63,04	0,64	63,64	0,81	80,85
Kreide	0,96	95,65	0,57	56,52	0,66	65,91	0,85	85,11
Schulter	0,96	95,65	0,54	54,35	0,66	65,91	0,85	85,11
Birke	0,8	80,43	0,46	45,65	0,48	47,73	0,85	85,11
Felder	0,91	91,3	0,54	54,35	0,64	63,64	0,85	85,11
Kirche	0,83	82,61	0,5	50	0,48	47,73	0,7	70,21
Probe	0,89	89,13	0,54	54,35	0,61	61,36	0,83	82,98
Gipfel	0,93	93,48	0,59	58,7	0,68	68,18	0,91	91,49
Brote	0,83	82,61	0,63	63,04	0,59	59,09	0,89	89,36
Türme	0,8	80,43	0,59	58,7	0,43	43,18	0,87	87,23
Freude	0,39	39,13	0,5	50	0,41	40,91	0,64	63,83
Kerze	0,85	84,78	0,39	39,13	0,55	54,55	0,77	76,6
Scherben	0,63	63,04	0,57	56,52	0,61	61,36	0,81	80,85
Puls	0,54	54,35	0,46	45,65	0,48	47,73	0,64	63,83
Konto	0,76	76,09	0,46	45,65	0,61	61,36	0,81	80,85
Graben	0,7	69,57	0,63	63,04	0,5	50	0,83	82,98
plus	0,67	67,39	0,59	58,7	0,57	56,82	0,77	76,6
Flasche	0,63	63,04	0,61	60,87	0,68	68,18	0,72	72,34
Wurm	0,67	67,39	0,65	65,22	0,7	70,45	0,85	85,11
Traum	0,63	63,04	0,65	65,22	0,61	61,36	0,85	85,11
Kreuze	0,57	56,52	0,43	43,48	0,55	54,55	0,77	76,6
Blase	0,54	54,35	0,59	58,7	0,52	52,27	0,79	78,72
Kragen	0,52	52,17	0,54	54,35	0,52	52,27	0,72	72,34
Würste	0,46	45,65	0,54	54,35	0,68	68,18	0,74	74,47
Torte	0,43	43,48	0,65	65,22	0,61	61,36	0,68	68,09
Glas	0,43	43,48	0,48	47,83	0,52	52,27	0,72	72,34
Knoten	0,37	36,96	0,63	63,04	0,48	47,73	0,66	65,96
Brause	0,39	39,13	0,65	65,22	0,64	63,64	0,7	70,21
Kirsche	0,37	36,96	0,43	43,48	0,55	54,55	0,64	63,83
Korken	0,3	30,43	0,5	50	0,34	34,09	0,57	57,45
Frost	0,22	21,74	0,59	58,7	0,66	65,91	0,53	53,19
Gurke	0,24	23,91	0,57	56,52	0,64	63,64	0,57	57,45
Schalter	0,22	21,74	0,54	54,35	0,7	70,45	0,57	57,45
falsche	0,15	15,22	0,28	28,26	0,25	25	0,32	31,91
Krone	0,2	19,57	0,5	50	0,57	56,82	0,53	53,19
Blume	0,11	10,87	0,57	56,52	0,66	65,91	0,47	46,81
Turm	0,11	10,87	0,48	47,83	0,75	75	0,47	46,81
Kreise	0,09	8,7	0,52	52,17	0,61	61,36	0,4	40,43
Birne	0,07	6,52	0,52	52,17	0,59	59,09	0,43	42,55
Frau	0,07	6,52	0,48	47,83	0,64	63,64	0,49	48,94
Karte	0,07	6,52	0,54	54,35	0,57	56,82	0,34	34,04
Wolken	0,07	6,52	0,7	69,57	0,66	65,91	0,43	42,55
Bruder	0,07	6,52	0,5	50	0,61	61,36	0,36	36,17
Turner	0,04	4,35	0,46	45,65	0,64	63,64	0,43	42,55
Frucht	0,04	4,35	0,46	45,65	0,48	47,73	0,36	36,17
Galopp	0,04	4,35	0,57	56,52	0,5	50	0,36	36,17
Würmer	0,02	2,17	0,57	56,52	0,61	61,36	0,36	36,17
Kinder	0,02	2,17	0,43	43,48	0,59	59,09	0,43	42,55
Gurken	0,02	2,17	0,52	52,17	0,73	72,73	0,4	40,43
Bluse	0,02	2,17	0,46	45,65	0,66	65,91	0,36	36,17
Wurst	0,02	2,17	0,63	63,04	0,7	70,45	0,3	29,79

Furcht	0,02	2,17	0,26	26,09	0,27	27,27	0,19	19,15
Kunst	0,02	2,17	0,54	54,35	0,48	47,73	0,34	34,04
Gelder	0,02	2,17	0,54	54,35	0,57	56,82	0,36	36,17
Mittelwert	0,46	45,94	0,53	53,06	0,58	57,68	0,64	63,59

Jahrgangsstufe 4

	MZP 1		MZP 2		MZP 3		MZP 4	
Items	Mittelwert	P_i	Mittelwert	P_i	Mittelwert	P_i	Mittelwert	P_i
Gleis	1	100	0,79	78,95	0,8	80	0,92	91,89
schreiben	1	100	0,74	73,68	0,8	80	0,78	78,38
Wolke	1	100	0,87	86,84	0,85	85	0,95	94,59
Blüte	1	100	0,68	68,42	0,85	85	0,86	86,49
Rinder	0,95	94,87	0,74	73,68	0,83	82,5	0,92	91,89
Forst	0,79	79,49	0,58	57,89	0,68	67,5	0,78	78,38
Traube	1	100	0,82	81,58	0,83	82,5	0,81	81,08
Durst	0,95	94,87	0,79	78,95	0,83	82,5	0,89	89,19
Kreide	1	100	0,68	68,42	0,8	80	0,78	78,38
Schulter	0,97	97,44	0,74	73,68	0,8	80	0,86	86,49
Birke	0,97	97,44	0,68	68,42	0,83	82,5	0,86	86,49
Felder	0,95	94,87	0,47	47,37	0,68	67,5	0,78	78,38
Kirche	0,87	87,18	0,74	73,68	0,7	70	0,84	83,78
Probe	0,97	97,44	0,79	78,95	0,75	75	0,97	97,30
Gipfel	0,97	97,44	0,74	73,68	0,75	75	0,86	86,49
Brote	0,97	97,44	0,71	71,05	0,78	77,5	0,73	72,97
Türme	0,97	97,44	0,82	81,58	0,83	82,5	0,84	83,78
Freude	0,62	61,54	0,68	68,42	0,73	72,5	0,68	67,57
Kerze	0,97	97,44	0,71	71,05	0,75	75	0,84	83,78
Scherben	0,77	76,92	0,66	65,79	0,8	80	0,84	83,78
Puls	0,67	66,67	0,47	47,37	0,58	57,5	0,57	56,76
Konto	0,95	94,87	0,68	68,42	0,823	82,5	0,73	72,97
Graben	0,95	94,87	0,74	73,68	0,9	90	0,92	91,89
plus	0,95	94,87	0,74	73,68	0,7	70	0,78	78,38
Flasche	0,9	89,74	0,71	71,05	0,83	82,5	0,78	78,38
Wurm	0,95	94,87	0,63	63,16	0,85	85	0,81	81,08
Traum	0,92	92,31	0,68	68,42	0,83	82,5	0,89	89,19
Kreuze	0,95	94,87	0,71	71,05	0,8	80	0,84	83,78
Blase	0,95	94,87	0,71	71,05	0,78	77,5	0,76	75,68
Kragen	0,9	89,74	0,58	57,89	0,83	82,5	0,89	89,19
Würste	0,79	79,49	0,61	60,53	0,75	75	0,92	91,89
Torte	0,82	82,05	0,74	73,68	0,78	77,5	0,73	72,97
Glas	0,82	82,05	0,68	68,42	0,83	82,5	0,89	89,19
Knoten	0,79	79,49	0,61	60,53	0,7	70	0,84	83,78
Brause	0,82	82,05	0,68	68,42	0,88	87,5	0,76	75,68
Kirsche	0,77	76,92	0,63	63,16	0,7	70	0,78	78,38
Korken	0,79	79,49	0,61	60,53	0,7	70	0,70	70,27
Frost	0,69	69,23	0,55	55,26	0,78	77,5	0,78	78,38
Gurke	0,77	76,92	0,74	73,68	0,68	67,5	0,81	81,08
Schalter	0,74	74,36	0,63	63,16	0,73	72,5	0,89	89,19
falsche	0,51	51,28	0,42	42,11	0,55	55	0,59	59,46
Krone	0,64	64,1	0,79	78,95	0,8	80	0,84	83,78
Blume	0,56	56,41	0,71	71,05	0,68	67,5	0,86	86,49
Turm	0,54	53,85	0,74	73,68	0,83	82,5	0,86	86,49
Kreise	0,44	43,59	0,76	76,32	0,85	85	0,70	70,27
Birne	0,41	41,03	0,63	63,16	0,8	80	0,73	72,97
Frau	0,33	33,33	0,76	76,32	0,83	82,5	0,89	89,19
Karte	0,33	33,33	0,61	60,53	0,73	72,5	0,81	81,08
Wolken	0,31	30,77	0,82	81,58	0,78	77,5	0,89	89,19
Bruder	0,28	28,21	0,71	71,05	0,85	85	0,81	81,08
Turner	0,23	23,08	0,66	65,79	0,75	75	0,70	70,27
Frucht	0,21	20,51	0,68	68,42	0,75	75	0,81	81,08
Galopp	0,18	17,95	0,68	68,42	0,9	90	0,86	86,49
Würmer	0,15	15,38	0,55	55,26	0,9	90	0,97	97,30
Kinder	0,15	15,38	0,76	76,32	0,83	82,5	0,81	81,08
Gurken	0,13	12,82	0,68	68,42	0,65	65	0,81	81,08
Bluse	0,1	10,26	0,71	71,05	0,73	72,5	0,73	72,97
Wurst	0,08	7,69	0,71	71,05	0,88	87,5	0,92	91,89
Furcht	0,08	7,69	0,37	36,84	0,38	37,5	0,49	48,65
Kunst	0,08	7,69	0,71	71,05	0,8	80	0,76	75,68
Gelder	0,05	5,13	0,79	78,95	0,68	67,5	0,89	89,19
Mittelwert	0,68	67,89	0,76	75,54	0,77	76,93	0,81	81,39