Europa-Universität Flensburg

Institut für Sonderpädagogik

Sprachsystematische Lernverlaufsdiagnostik zur differenzierten Erfassung von Rechtschreibkompetenz mit Levumi - Konstruktion, Evaluation und Implementation des webbasierten Rechtschreibkompetenz-Messverfahrens (ReKoMe)

Dissertation

zur Erlangung des akademischen Grades Doktor der Philosophie (Dr. phil.)

Erstgutachterin: Prof.in Dr. Marie-Christine Vierbuchen

Zweitgutachter: Prof. Dr. Andreas Mühling

Vorgelegt von:

Lisa Mau
geb. am 21.01.1986 in Eckernförde

Tag der Abgabe:
03.04.2023

Vorwort

Das Dissertationsprojekt wurde am Institut für Sonderpädagogik an der Europa-Universität Flensburg und am Institut für Informatik der Christian-Albrechts-Universität zu Kiel durchgeführt und ist Teil des interdisziplinären Projekts Levumi. Die Entwicklung von Instrumenten zur Lernverlaufsdiagnostik im Bereich Rechtschreibung erfordert eine neue theoretische Fundierung, die auf den tatsächlichen individuellen Schriftentwicklungen der Schüler*innen und sprachsystematischen Erkenntnissen basiert. Die Autorin hat im Rahmen ihrer Forschung den Bereich Rechtschreibung in die Onlinelernplattform Levumi integriert und dazu das Rechtschreibkompetenz-Messverfahren (ReKoMe)[1] zur automatisierten sprachsystematischen Lernverlaufsdiagnostik entwickelt, evaluiert und implementiert. Dies ist bislang ein Novum. Das ReKoMe ist eine Erweiterung des bestehenden Instrumentariums zur Lernverlaufsdiagnostik im Bereich Rechtschreibung, das Praktiker*innen zur Verfügung steht, um Grundschüler*innen bei der Entwicklung ihrer Rechtschreibkompetenz bestmöglich zu unterstützen.

Für den erfolgreichen Abschluss meines Dissertationsprojektes danke ich zuallererst ganz besonders meiner Doktormutter Frau Professorin Dr. Marie-Christine Vierbuchen für die anregende und wertschätzende Begleitung meines Arbeitsprozesses. Mein ganz besonderer Dank gilt auch Herrn Professor Dr. Andreas Mühling, der mich fachlich und persönlich unterstützend durch dieses Dissertationsprojekt begleitet hat. Als Wegbegleiter reichte seine Rolle weit über die Funktion des Zweitbetreuers hinaus. Ich danke für seine Bereitschaft, meine Arbeit im Hinblick auf die empirischen Herausforderungen zu unterstützen und für die wertvollen Anregungen, die er mir für die Umsetzung gab. Ein herzlicher Dank gilt auch Frau Professorin Dr. Inge Blatt, die hinsichtlich theoretischer und praxisbezogener Fragen immer ein offenes Ohr für mich hatte und mir Hilfestellung leistete. Ich danke allen beteiligten Schüler*innen, Lehrer*innen, Studierenden und wissenschaftlichen Hilfskräften. Ohne ihre Unterstützung wäre die Studie nicht möglich gewesen. Mein größter Dank gilt abschließend meiner Familie und meinen Freundinnen, die mich auf meinem Weg unterstützt und begleitet haben.

[1] Das Rechtschreibkompetenz-Messverfahren (ReKoMe) ist seit 2017 auf der Onlinelernplattform www.Levumi.de bisher unter dem Namen „Wortdiktat“ (Mau, 2017) implementiert.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

DIF Differential Item Functioning

HSP Hamburger Schreib-Probe

IRT Item-Response-Theorie

KTT Klassische Test Theorie

Lup Lupenstellen

LA Lupenstellen Anzahl

Levumi Lernverlaufsmonitoring

LU Lupenstellen Ursprungsitempool

Morph Morphologisches Prinzip

Peri Peripheriebereich

PhonSilb Phonographisch - Silbisches Prinzip

Quan Quantitativ

ReKoMe Rechtschreibkompetenz-Messverfahren

ReKoMe-PP Rechtschreibkompetenz-Messverfahren Paper-Pencil

Wortbil Prinzip der Wortbildung

1 Einleitung

Das Erlernen der Grundfertigkeiten des Schreibens gilt im Primarbereich als eine in den Bildungsstandards für den Deutschunterricht normativ gesetzte Kompetenzerwartung (Kultusminister Konferenz, 2005), die zum Ende der Grundschulzeit von allen Schüler*innen als Regelstandard erreicht werden soll. Die Fähigkeit, orthografisch normgerecht schreiben zu können, beeinflusst den Lernerfolg in allen Schulfächern, ist zentral für den weiteren Bildungsweg eines Kindes und gilt als eine grundlegende Kulturtechnik, die erst zur gesellschaftlichen Teilhabe in der heutigen Wissensgesellschaft befähigt. Für den erfolgreichen Übergang in die weiterführende Schule wird der Rechtschreibkompetenz sogar eine größere Bedeutung beigemessen als der Intelligenz (Schneider, 2008). Für eine frühzeitige und engmaschige Begleitung der individuellen Lernverläufe im Bereich Rechtschreibung sprechen die Studienergebnisse der empirischen Bildungsforschung zur Rechtschreibkompetenz deutscher Grundschüler*innen, sowie die Ergebnisse aus der Lehr-Lernforschung zum domänenspezifischen Wissen von Lehrkräften und deren diagnostischen Kompetenzen im Bereich Orthografie.

Die Befunde der empirischen Bildungsforschung zeigen seit langem, dass das Thema Rechtschreibung von Anfang an mit vielfältigen Ausgangslagen und Problemen verbunden ist und dass es eine erhebliche Diskrepanz zwischen den in den Bildungsstandards verankerten Kompetenzerwartungen und den tatsächlich erreichten Leistungen von Grundschüler*innen gibt. Der nationale IQB-Bildungstrend hat 2021 zum dritten Mal überprüft, inwieweit die schulischen Leistungen in den Fächern Deutsch (Lesen, Zuhören, Orthografie) und Mathematik den in den jeweiligen Bildungsstandards (Kultusminister Konferenz, 2005) zugrunde gelegten Kompetenzerwartungen entsprechen (Stanat et al., 2022). Die Ergebnisse sind besorgniserregend. Der Leistungsabfall im Bereich Orthografie bei Grundschüler*innen der vierten Klasse entspricht einem Viertel Schuljahr (Stanat et al., 2022). Der Anteil der Schüler*innen, die die Mindeststandards nicht erreichen, steigt deutlich von 22% auf 30%, warnt die Bildungsforscherin Stanat:

> „Diese Zahlen müssen uns alarmieren, denn für diese Schüler*innen steht die weitere schulische Laufbahn auf dem Spiel. [...] Die Lage ist wirklich besorgniserregend.[...] Wir brauchen vor allem eine langfristig angelegte gezielte Förderung von Kindern, die in Gefahr sind, die Mindeststandards nicht zu erreichen. Dabei müssen zunächst die Basiskompetenzen in Deutsch und Mathematik gesichert werden [...] ohne sie haben die Kinder kaum eine Chance“(Kerstan, 2022, 1. Juli).

Obwohl die diagnostische Tätigkeit und die Begleitung der Lernprozesse der Schüler*innen zu den Kernkompetenzen einer Lehrkraft gehören (Kultusminister Konferenz, 2005), ver-

fügen dies nicht immer in ausreichendem Maße über fachliche und diagnostische Kompetenzen im Bereich des Schriftspracherwerbs, um die Lernprozesse ihrer Schüler*innen adäquat begleiten und deren Lern- und Leistungsstand richtig einschätzen zu können (Corvacho del Toro & Günther, 2013; Schröder, 2019). Der erfolgreiche Schriftspracherwerb steht und fällt somit auch mit der jeweiligen Kompetenz der Lehrkraft (Roos & Schöler, 2009).

Die Heterogenität der Lernausgangslagen und Lernprozesse beim Erwerb der Rechtschreibkompetenz sowie die diagnostischen Fähigkeiten der Lehrkräfte erfordern eine regelmäßige Lernstandserhebung mit standardisierten und reliablen Testverfahren. Das Konzept der Lernverlaufsdiagnostik gilt vor diesem Hintergrund als das wichtigste Instrument für erfolgreiche Lernprozesse und zur Verbesserung von Bildungsqualität. Ziele der Lernverlaufsdiagnostik sind die schulische Prävention, die formative Evaluation des Unterrichts und die Grundlage für datenbasierte Förderentscheidungen (Blumenthal, 2022). Mit dem Einsatz von Instrumenten zur Lernverlaufsdiagnostik sind eine Vielzahl an positiven Effekten verbunden, wie z.B. ein höherer Lernzuwachs und eine Verbesserung der Unterrichtsqualität. Entgegen der Bedeutung der individuellen Lernverlaufsdiagnostik für gelingende Lernprozesse im Bereich Rechtschreibung ist die Umsetzung des Konzepts im deutschsprachigen Raum nicht zufriedenstellend realisiert.

Die schriftsprachtheoretische Fundierung eines Instruments zur differenzierten Lernverlaufsdiagnostik ist zentral, da auf dieser Grundlage die Auswahl und Zuordnung von Kategorien zur Analyse von Schreibprodukten erfolgt, die individuelle Rechtschreibkompetenz differenziert beschrieben werden kann und sich daraus unmittelbare Konsequenzen für anschließende Fördermaßnahmen ergeben (Naujokat, 2015). In der Rechtschreibforschung und Fachdidaktik wird vor dem Hintergrund der Ergebnisse großer Schulleistungsstudien (z.B. Pisa, NEPS) zur Modellierung und Entwicklung von Rechtschreibkompetenz seit langem Kritik an einem normorientierten Verständnis des Schriftspracherwerbs geübt (z.B. Blatt et al., 2021 Hinney, 2010; Jagemann & Weinhold, 2018; Weinhold et al., 2020) und einen Paradigmenwechsel hin zu einer sprachsystematischen Modellierung des Konstrukts Rechtschreibung gefordert (Blatt, Prosch & Lorenz, 2016). Die Entwicklung von Instrumenten zur Lernverlaufsdiagnostik erfordert eine neue theoretische Fundierung, die auf den tatsächlichen individuellen Schriftentwicklungen der Schüler*innen und sprachsystematischen Erkenntnissen basiert. Der Diskussionsstand zeigt, dass die bisher in bestehenden Instrumenten zur Lernverlaufsdiagnostik verwendeten normbasierten Stufenmodelle des Schriftspracherwerbs, in denen das lautorientierte Schreiben als notwendiger Entwicklungsschritt auf dem Weg zum orthografischen Schreiben angesehen wird, den heterogenen Lernausgangslagen und Lernverläufen nicht gerecht werden können. Demgegenüber eröffnet das struktur- und prozessorientierte, sprachsystematische Rechtschreibkompetenzmodell große Potenziale für eine individualisierte Lernverlaufsdiagnostik im Spannungsfeld zwischen generalisierenden Modellen und individualisierter Betrachtung der Rechtschreibkompetenzentwicklung (Bulut, 2018). Das Modell berücksichtigt die sprachlichen Voraussetzungen der Schüler*innen und setzt anstelle des lautgetreuen Schreibens oder des Regellernens das Erkunden und Verstehen von Schriftstrukturen für den Erwerb von Rechtschreibkompetenz (Blatt et al., 2015). Damit sind große Potenziale für die Lernverlaufsdiagnostik verbunden, weil eine schriftsystematische Analyse von

Schreibprodukten möglich ist, welche die tatsächlichen individuellen Schriftentwicklungen der Schüler*innen „fernab der tradierten Brille einer idealtypischen Entwicklung“ überprüft (Weinhold et al., 2020, S.28).

Das wissenschaftliche Erkenntnisinteresse dieser Arbeit beruht auf zwei Forschungsdesideraten und hat einen explorativen Charakter: Eine sprachsystematische Fundierung bei der Konstruktion von Instrumenten zur Lernverlaufsdiagnostik wird in der empirischen Sonderpädagogik bislang nicht berücksichtigt. Damit bleiben wichtige Potenziale für den Rechtschreibunterricht, für Diagnostik und Förderung sowie für den Erkenntnisgewinn in Bildungsforschung und Fachdidaktik über individuelle Rechtschreibkompetenzentwicklungen ungenutzt. Zweitens fehlt bisher ein für Praxis und Forschung zugänglicher Algorithmus, der eine automatisierte, theoretisch fundierte qualitative Rechtschreibanalyse auf der Basis des sprachsystematischen Rechtschreibkompetenzmodells ermöglicht. Dies ist vor allem aus einer praxisrelevanten Sichtweise wichtig, da die Auswertung von Schreiblösungen auf der Basis dieses Modells Fachwissen erfordert und zeitaufwendig und komplex ist (Frahm, 2013). Darüber hinaus ist eine differenzierte, theoretisch fundierte Testauswertung von entscheidender Bedeutung, die über das bloße Auszählen von richtig und falsch geschriebenen Wörtern hinausgeht und die Art der individuellen Schreiblösungen in den Mittelpunkt stellt. Probleme von Schüler*innen in der Rechtschreibung können nämlich in deutlich abgrenzbaren Bereichen liegen, auch wenn die quantitativen Ausprägungen gleich sind (Corvacho del Toro, 2016). Ausgehend von dem beschriebenen Problemaufriss lauten die zwei zentralen Forschungsfragen der Arbeit:

1. Inwiefern kann auf Basis des sprachsystematischen Rechtschreibkompetenzmodells ein webbasiertes, zeit- und testökonomisches Instrument zur differenzierten Lernverlaufsdiagnostik im Primarbereich entwickelt werden?

Das Instrument soll Lehrkräfte dabei unterstützen, die individuellen Lernstände und -verläufe ihrer Schüler*innen beim Rechtschreibkompetenzerwerb zuverlässig zu diagnostizieren, engmaschig zu begleiten und Lernlücken frühzeitig zu identifizieren, um passende Fördermaßnahmen einleiten zu können.

2. Inwiefern kann ein Algorithmus entwickelt werden, der die Testergebnisse auf Ganzwortebene und auf der Ebene orthografischer Teilkompetenzen automatisiert analysiert, codiert und zuverlässige Ergebnisse liefert?

Für eine erfolgreiche Entwicklung der Rechtschreibkompetenz ist ein sprachsystematischer, am individuellen Lernstand orientierter Rechtschreibunterricht wesentlich. Das Verständnis der Systematik der Rechtschreibung ist dafür die Grundvoraussetzung. Mit dem Algorithmus soll die Lernentwicklung der Schüler*innen differenziert und zuverlässig eingeschätzt werden, um den Einfluss des Rechtschreibunterrichts auf die individuelle Kompetenzentwicklung zu beurteilen zu können. Die automatisierte Testauswertung und Ergebnisdarstellung bildet die Grundlage für die Konzeption individueller Förder- und Lernangebote.

Gegenstand der vorliegenden empirischen Studie ist die Konstruktion, Pilotierung, Evaluation und Implementation des webbasierten Rechtschreibkompetenz-Messverfahrens (Re-

KoMe) [1] einschließlich der Entwicklung eines Algorithmus zur automatisierten differenzierten Lernverlaufsdiagnostik, das test- und zeitökonomisch im Unterricht eingesetzt werden kann und die besonderen Potenziale der sprachsystematischen Sichtweise auf den Schriftspracherwerb berücksichtigt. Für den Einsatz von Instrumenten zur Lernverlaufsdiagnostik im Schulalltag ist eine ökonomische Durchführung und automatisierte Auswertung der Testergebnisse vor dem Hintergrund der Praktikabilität und der Akzeptanz Grundvoraussetzung. Dies kann durch eine webbasierte Diagnostik umgesetzt werden. Die gewonnenen Informationen über Lernstände und Lernverläufe werden durch die Implementierung auf der Onlineplattform Levumi für Akteur*innen im Bildungswesen nutzbar gemacht und liefern der Bildungsforschung und Fachdidaktik wichtige Erkenntnisse über die Entwicklung der Rechtschreibkompetenz von Grundschüler*innen.

Die Arbeit ist in einem interdisziplinären Forschungsfeld angesiedelt, zu dem unterschiedliche Disziplinen wie Linguistik, Spracherwerbsforschung, Fachdidaktik Deutsch sowie Psychologie, Grundschulpädagogik, Inklusionspädagogik und Informatik gehören. Das ReKoMe schließt eine bestehende Forschungslücke und dient sowohl der Grundlagenforschung als auch der praxisorientierten Anwendung. Es ist ein zentraler Baustein einer umfassenden unterrichtsimmanenten diagnostischen Gesamtstrategie im Kontext einer adaptiven und teil-digitalisierten Unterrichtsarchitektur. Der Einsatz von ReKoMe trägt entscheidend dazu bei, die Validität pädagogischer Entscheidungen im Unterricht zum Schriftspracherwerb zu erhöhen. Dies verbessert die Unterrichtsqualität und die Bildungschancen aller Schüler*innen.

Die Arbeit ist wie folgt gegliedert: Im *theoretischen Teil* werden zunächst in Kapitel 2 die Kontroversen innerhalb der Fachwissenschaft und -didaktik zu Rechtschreibkonzepten (Dependenz-, Autonomie- und Interdependenzhypothese), zur Stellung der graphematischen Ebene im Sprachsystem und zum Verhältnis von gesprochener und geschriebener Sprache sowie Forschungsergebnisse zur Graphematik nach Eisenberg dargestellt. Diese theoretische Auseinandersetzung ist notwendig, da unterschiedliche (lern-)psychologische und fachliche Zugänge zur Domäne Rechtschreibung zu unterschiedlichen Positionen führen und damit auch unterschiedliche theoretische Konsequenzen für die Konstruktion des ReKoMe nach sich ziehen. In Kapitel 3 werden die Potenziale einer sprachsystematischen Modellierung von Rechtschreibkompetenz als Grundlage für die Konstruktion eines Instruments zur Lernverlaufsdiagnostik und deren Implikationen für den Schriftspracherwerbsunterricht erschlossen. Darüber hinaus thematisiert das Kapitel die Bildungsstandards für das Fach Deutsch in der Primarstufe und die Diskussion um den Kompetenzbegriff in der empirischen Bildungsforschung und beschreibt die Anforderungen an die Modellierung und Messung von Rechtschreibkompetenz, da diese Aspekte bei der Konstruktion eines Instruments zur kompetenzorientierten Lernverlaufsdiagnostik eine Rolle spielen. In Kapitel 4 folgt die Betrachtung des Konzepts der Lernverlaufsdiagnostik mit seinen Zielen, Methoden und Gelingensbedingungen. Daran schließt sich eine Analyse der bisher vorliegenden Instrumente zur Lernverlaufsdiagnostik im Bereich Rechtschreibung an. Das Kapitel diskutiert auch die besonderen methodischen Herausforderungen, die bei der Entwicklung von Instrumenten zur Lernverlaufsdiagnostik im Bereich Rechtschreibung verbunden sind.

[1] Das Rechtschreibkompetenz-Messverfahren (ReKoMe) ist seit 2017 auf der Onlinelernplattform www.Levumi.de bisher unter dem Namen „Wortdiktat“ (Mau, 2017) implementiert.

Die Zusammenfassung der theoretischen Erkenntnisse sowie die Darstellung der Ziele und Fragestellungen sind Gegenstand von Kapitel 5. Der Konstruktionsprozess des ReKoMe, die digitale Umsetzung und Implementierung einschließlich der Entwicklung des Algorithmus zur automatisierten Ergebnisanalyse, die Entwicklung einer Tastaturschulung sowie die zusammenfassende Darstellung der konstruierten Instrumente werden im 6. Kapitel beschrieben.

Der *empirische Teil* liefert einleitend mit Kapitel 7 eine Beschreibung des methodischen Vorgehens zur Pilotierung und Evaluierung des konstruierten ReKoMe und wird ausgehend vom Untersuchungsplan dargestellt. Mittels statistischer Analysen der klassischen Testtheorie und der Item-Response-Theorie erfolgt die Pilotierung und Evaluation des Paper-Pencil Tests, des webbasierten Prototyps, des Algorithmus und des Messverfahrens. Grundlage der Studie ist eine umfangreiche Stichprobe von 312 Schüler*innen aus 17 Grundschulklassen in drei Bundesländern. Insgesamt liegen für die Evaluationsstudie durchschnittlich 5135 beantwortete Testaufgaben vor. Das Kapitel beschreibt die Analysemethoden zur Evaluierung, wobei die Verfahren der Klassischen Testtheorie und der Item-Response-Theorie im Fokus stehen. Die Anwendung dieser Analyseverfahren ermöglicht eine fundierte Aussage über die Qualität des Messverfahrens. Im Kapitel 8 steht die Pilotierung im Fokus. Ziel der Pilotstudien ist es, die Qualität der konstruierten Testaufgaben, des Algorithmus und des webbasierten Testdesigns unter Alltagsbedingungen im System Schule zu überprüfen. Die Ergebnisse der Studien sollen auch dazu beitragen, die Integrierbarkeit des Messverfahrens in den Unterricht zu überprüfen und Hinweise auf mögliche Verbesserungen zu identifizieren. Die Ergebnisse der Paper-Pencil Studie dienen als Datengrundlage für die Entwicklung und Überprüfung des Algorithmus. Im Anschluss stehen die Ergebnisse der Prototypenstudie im Fokus. In Kapitel 9 erfolgt die empirische Überprüfung der psychometrischen Güte des konstruierten Messverfahrens anhand von Analysen der klassischen Testtheorie und der Item-Response-Theorie sowie die Überprüfung der faktoriellen Struktur des zugrunde liegenden theoretischen Rahmenmodells. Die Ergebnisse der Evaluationsstudie werden im Kontext der Beschreibung der Stichprobenkonstruktion, der Untersuchungsdurchführung und der angewandten statistischen Analysen vorgestellt. Daran schließt die Beantwortung der Forschungsfragen der Arbeit und die Darstellung von Einzelfallbeschreibungen an, um das Potenzial des Messverfahrens zur Steigerung der pädagogischen Validität und zum Abbau von Bildungsdisparitäten zu verdeutlichen. Abschließend werden die Ergebnisse im Hinblick auf die in der Zielsetzung und der Fragestellung formulierten Forschungsfragen in Kapitel 10 diskutiert. Es folgt eine Darstellung der Grenzen der Studie und ein Ausblick auf zukünftige Forschung.

2 Sprachsystematische Grundlagen

Innerhalb der fachdidaktischen Forschung zum Schrift- und Orthografieerwerb werden die schriftsystematischen Grundlagen und die damit verbundenen didaktischen Ansätze kontrovers diskutiert (Kruse & Reichardt, 2016). Es besteht derzeit kein Konsens darüber, wie Rechtschreibkompetenz zu modellieren und zu operationalisieren ist. Vielmehr werden unterschiedliche Standpunkte vertreten und verschiedene Erwerbsmodelle zum Schriftspracherwerb als Ausgangspunkt für die Konzeptualisierungen von Rechtschreibkompetenz gewählt (Hinney, 1997; Naujokat, 2015). Diese Auseinandersetzung mit den divergierenden Ansichten wird aufgrund unterschiedlicher (Lern-)psychologischer und fachlicher Zugänge zur Domäne Rechtschreibung notwendig, die zu verschiedenen Positionen führen und somit auch unterschiedliche theoretische Konsequenzen für die Konstruktion des Rechtschreibkompetenz-Messverfahrens nach sich ziehen. Zentraler Ausgangspunkt ist es, solche vorhandenen Theorien in den Blick zu nehmen und im Kontext der Forschungsfrage einzuordnen. Im Kapitel 2.1 wird die Kontroverse innerhalb der Fachwissenschaft und Didaktik zu Orthografiekonzepten (Dependenz-, Autonomie- und Interdependenzhypothese), zur Stellung der graphematischen Ebene im Sprachsystem und zum Verhältnis von gesprochener und geschriebener Sprache (vgl. Kap. 2.1.3) beschrieben.

Die Graphematik bietet einen neuen Blickwinkel auf die Rechtschreibung und eröffnet damit neue Möglichkeiten für die Didaktik, Diagnostik und Förderung von Rechtschreibkompetenz. Die Erkenntnisse der Graphematik nach Eisenberg werden im Kapitel 2.2 und ihre didaktische Umsetzung im Kapitel 2.2.5 dargestellt.

2.1 Graphematik und Orthografie

Im Folgenden wird das Forschungsgebiet der Graphematik, das sich mit dem Schriftsystem und der Normierung der Schreibung (Orthografie) beschäftigt, beschrieben. Im Deutschen gibt es nur selten Abweichungen von einem graphematisch rekonstruierbaren Schriftsystem. Dies bietet wichtige Vorteile für den schriftsprachlichen Anfangsunterricht, da auf der Basis einfacher graphematischer Regelmäßigkeiten richtige Wortschreibungen abgeleitet werden können.

Eine wesentliche Grundlage für die Rechtschreibung ist eine orthografisch und syntaktisch korrekte Verschriftlichung. Das hierfür relevante Forschungsgebiet stellt die Graphematik dar. Die Graphematik beschreibt das Schriftsystem, die Orthografie, als ein Teilgebiet der Graphematik, beschäftigt sich mit der Normierung des Schriftsystems. Während die Graphematik „einen Lösungsraum möglicher Schreibungen“ (Neef, 2011, S.11) für Lautungen

vorgibt, die ein Wort repräsentieren können, legt die Orthografie fest, welche Schreibung unter allen möglichen als korrekt gilt (Dürscheid, 2016). Die Graphematik stellt zudem die der normalen Schreibung zugrunde liegenden Regularitäten heraus (Eisenberg, 2020a).

Eisenberg (2020a) hebt die Bedeutung einer normierten Schreibung als etwas funktional Erforderliches hervor, das nicht als erzwungen, sondern als ein Ergebnis eines natürlich ablaufenden Prozesses anzusehen und für die Einheitlichkeit und Stabilität von Formen für eine Sprache Voraussetzung ist.

Innerhalb der orthografischen Norm des Deutschen gibt es nur in seltenen Fällen Abweichungen von einem graphematisch rekonstruierbaren Schriftsystem. So lassen sich auf der Basis einfacher graphematischer Regularitäten korrekte Wortschreibungen ableiten (Eisenberg, 2020a).

2.1.1 Historische Entwicklung

Seit den Beschlüssen der 2. Orthografischen Konferenz 1901 existiert im Deutschen eine explizit normierte staatlich verordnete Orthografie, die in einem amtlichen Regelwerk festgelegt ist (Rat für deutsche Rechtschreibung, 2018). Sie beziehen sich auf Laut-Buchstaben-Zuordnungen, Getrennt- und Zusammenschreibung, Groß- und Kleinschreibung, Zeichensetzung sowie Worttrennung am Zeilenende (Rat für deutsche Rechtschreibung, 2018).

Seit 1996 wurde diese im Zuge der neuen Rechtschreibreform verändert (Bußmann, 2008; Ossner, 2018). Im Unterschied zum amtlichen Regelwerk, das sich als weitgehend unsystematisch darstellt, systematisiert die Graphematik den Kernbereich der Rechtschreibung (Blatt, 2010; Blatt et al., 2021). Eisenberg (2020a, S.313) fasst pointiert den Sachverhalt wie folgt zusammen:

> „Wer eine Orthographie erwirbt, lernt nicht nur schreiben, sondern er lernt richtig im Sinne von normgerecht schreiben. Entscheidend ist letztlich, wie das geschriebene Wort aussieht. Unwichtig ist, nach welchen Regeln die Schreibung zustande kommt. Ein Orthographiefehler ist vorhanden oder nicht vorhanden, soweit die jeweilige Schreibung in der Orthographie geregelt ist. Eine Graphematik ermittelt dagegen die Regularitäten, die dem normalen Schreiben zugrunde liegen.“

Orthografie stellt zwar ein normiertes und kodifiziertes Rechtschreibwissen zur Verfügung, das erlernt werden muss, ist jedoch für den Erwerb der Rechtschreibung wenig hilfreich (Reichardt, 2015, S. 51).

> „Im Gegensatz zur verbreiteten Auffassung von Rechtschreiblernen als einer Erfüllung gesellschaftlicher Normen richtet die Graphematik den Blick auf das hohe Lernpotenzial des Rechtschreiblernens für die Entwicklung der mündlichen und schriftlichen Sprachkompetenz (Eisenberg 2004, Eisenberg/Fuhrhop 2007)“ (Blatt, 2010, S.102).

Bereits 1974 ist das deutsche Schriftsystem von Riehme in historisch geprägte orthografische Prinzipien untergliedert worden. Die, die folgende Funktionen innerhalb des Sprachsystems einnehmen:

- phonologisches Prinzip: Phonem-Graphem-Korrespondenzen spielen bei der Verschriftlichung von Wörtern eine Rolle
- morphematisches Prinzip: Wörter werden nach dem Stammprinzip trotz unterschiedlicher phonematischer Struktur gleich verschriftlicht, wie z.B. bei Auslautverhärtungen
- grammatisches Prinzip: regelt die grammatische Struktur der Sprache (Groß und Kleinschreibung)
- semantisches Prinzip: gleichlautende Wörter mit verschiedenen Bedeutungen werden unterschiedlich verschriftlicht, wie z.B. Lärche und Lerche
- historisches Prinzip: viele Schreibungen entsprechen einem früheren Aussprachemodus und wurden der heutigen gesprochenen Sprache nicht angepasst
- graphisch-formales Prinzip: regelt das Schriftbild eines Wortes

2.1.2 Wissenschaftliche Kontroversen innerhalb der Fachdisziplinen

Hinsichtlich der Ausdifferenzierung dieser Prinzipien wird innerhalb der Fachwissenschaft und –didaktik kontrovers diskutiert. Dabei geht es um die Stellung der graphematischen Ebene im Sprachsystem, um das Verhältnis von gesprochener und geschriebener Sprache sowie um die Herleitung von orthografischen Phänomenen und deren Ausnahmeschreibungen (Fay & Berkling, 2013). Insbesondere auf Eisenbergs (2016) Ausdifferenzierung des Aufbaus der deutschen Grammatik im aktuellen Grammatik-Duden wird häufig zurückgegriffen. Mit den jeweils unterschiedlich zugrunde gelegten Orthografietheorien sind verschiedene Schlussfolgerungen für die Schulpraxis verbunden. Sie stellen den Ausgangspunkt für die Schreibdidaktik dar, die den Lernenden Einsichten in das Schriftsystem ermöglichen sollen (Fay & Berkling, 2013). Zudem existieren Schriftspracherwerbskonzepte wie z.B. das Konzept „Lesen durch Schreiben“ von Reichen, die auf eine orthografietheoretische Fundierung verzichten und aus diesem Grunde in der Kritik stehen (Fay & Berkling, 2013). Im Folgenden werden die unterschiedlichen Positionen vorgestellt.

2.1.3 Das Verhältnis von gesprochener und geschriebener Sprache

Für die Konzeptualisierung von Rechtschreibung und deren Erwerbsprozesse bildet das Verständnis des Zusammenhangs von gesprochener und geschriebener Sprache einen zentralen Ausgangspunkt.

Zum einen gibt es Vertreter (Bußmann, 2008; Glück & Rödel, 2016; Neef, 2011), welche die Graphematik als Gegenstück zur segmentalen Phonologie verstehen, deren Forschungsfokus auf der Betrachtung der Beziehungen von schriftlichen und phonologischen

Repräsentationen liegt (Phonem-Graphem Korrespondenzen). Im »Lexikon der Sprachwissenschaft« (Bußmann, 2008, S.246) wird dazu ausgeführt:

> „Graphemik [Auch: Graphematik]. Wissenschaft von den distinktiven Einheiten des Schriftsystems [...]. Bei Alphabetschriften basiert G. auf Grund der Korrelationen zwischen gesprochener und geschriebener Sprache weitgehend auf den Analysemethoden der Phonologie.(Dependenzhypothese)“

Zum anderen wird die Graphematik der linguistischen Disziplin zugeordnet (u.a. Dürscheid, 2016; Eisenberg, 2020b), die sich sowohl auf die segmentalen als auch die suprasegmentalen Einheiten des Schriftsystems bezieht, d. h. bei der Betrachtung des Schriftsystems werden auch die Morphem-, die Wort-, die Satz- und die Textebene miteinbezogen (Dürscheid, 2016). Eisenberg postulierte bereits 1989 diesen weiter gefassten Gegenstandsbereich der Graphematik:

> „Als Heuristik bietet sich an, eine Graphematik analog zur Phonologie aufzubauen. Es gäbe dann eine segmentale Graphematik analog zur segmentalen Phonologie, eine mit der Silbe befaßte [sic.] Graphematik analog zur Silbenphonologie und eine lexikalische (Morphographematik) analog zur lexikalischen Phonologie, und natürlich reicht die Graphematik auch in die Syntax hinein (Eisenberg, 1989, S.59-S.60).“

Die unterschiedlichen Positionen stimmen darin überein, dass sie die Graphematik der Orthografie gegenüberstellen (Neef, 2011).

2.1.4 Dependenz-, Autonomie- und Interdependenzhypothese

Die Dependenz- und die Autonomiehypothese werden nicht nur unter dem Aspekt des Verhältnisses von gesprochener und geschriebener Sprache, sondern auch auf schriftsystematischer Ebene betrachtet (Dürscheid, 2016).

Die Vertreter der Dependenzhypothese (z.B. Coulmas, 1981; Paul, 1880; Saussure, 1916) gehen von einer Abhängigkeit des Schriftsystems vom Lautsystem und des Graphems vom Phonem aus (Dürscheid, 2016). Angenommen wird eine 1:1 Zuordnung von Lauten und Buchstaben, wobei die Graphemebene der Phonemebene nachgeordnet ist und die Rechtschreibung als eine direkte Abbildung der phonetischen Struktur und Schrift als sekundäres Schriftsystem gilt (Dürscheid, 2016; Noack, 2001). Korrespondenzregeln bilden den Rahmen für den Umgang mit der Umwandlung von Phonemen in Grapheme und mit Fällen, die keine phonemische Entsprechung haben, wie z.B. das silbentrennende h im Wort <sehen> oder das Dehnungs-h im Wort <Stuhl>. Diese werden als Ausnahmen von der optimalen Struktur betrachtet (Dürscheid, 2016; Noack, 2001). Es werden die folgenden Aspekte genannt, die für eine Abhängigkeit der geschriebenen von der gesprochenen Sprache sprechen:

- „Linguistisches Argument: Die Schrift ist nichts anderes als eine Visualisierung von Sprache, als in Buchstaben umgesetzter Schall.

- Entwicklungspsychologisches Argument: Die Schrift wird sowohl phylo- als auch ontogenetisch später erworben als Sprache.
- Logisches Argument: Sprache existiert ohne Schrift, Schrift aber nicht ohne Sprache.
- Funktionales Argument: Gesprochene Sprache wird bei weitaus mehr Gelegenheiten eingesetzt als die geschriebene“ (Dürscheid, 2016, S. 36).

Das Lautsystem und das Schriftsystem werden als relativ autonome Bereiche betrachtet. Sie interagieren zwar, sind aber voneinander unabhängig. Schrift wird als eine eigenständige Realisierungsform von Sprache verstanden (Autonomiehypothese). Vertreter dieser Position (z.B. Eisenberg, 1989; Maas, 2013) plädieren dafür, die Schrift als einen eigenen Forschungsgegenstand in Abgrenzung zur gesprochenen Sprache zu betrachten, was jedoch nicht ausschließt, dass Korrespondenzen zwischen Graphemen und Phonemen vorhanden sein können (Dürscheid, 2016).

Für diese Sichtweise werden u.a. folgende Argumente angeführt:

- „Strukturelles Argument: Die Schrift besteht aus diskreten Einheiten, die gesprochene Sprache stellt ein Kontinuum dar.
- Logisches Argument: Lesen und Schreiben rekurrieren nicht notwendigerweise auf die gesprochene Sprache.
- Linguistisches Argument: Die Schrift ermöglicht es, in Distanz zum Untersuchungsgegenstand zu treten. Sie macht sprachliche Strukturen der genauen Beobachtung zugänglich.
- Kulturwissenschaftliches Argument: Die Schrift bewahrt vor dem Vergessen, sie hat eine »dokumentarische Funktion« (Köller 1988: 157).
- Auf das Medium bezogenes Argument: Die Schrift hat optisch-visuelle Eigenschaften, die auf die gesprochene Sprache zurückwirken“ (Dürscheid, 2016, S. 38).

Radikale Vertreter der Dependenzhypothese lehnen jeglichen Zusammenhang zwischen der geschriebenen und gesprochenen Sprache ab, womit sie „die Existenz eines beiden Ausdrucksformen gemeinsamen Sprachsystems [leugnen] und ... die gesprochene und geschriebene Sprachform ein und derselben Spr. als sich gegenseitig fremde Spr.“ ansehen (Glück & Rödel, 2016, S. 79). Diese Position wurde besonders von H. Paul und F. de Saussure und wichtigen Vertretern des Strukturalismus befürwortet. Sie wird jedoch weitgehend als theoretisch nicht gerechtfertigt eingestuft und in der neueren Forschung abgelehnt (Glück & Rödel, 2016). Eine vermittelnde dritte Position stellt die Interdependenzhypothese dar, die als eine abgeschwächte und relativierte Form der Autonomiehypothese gilt, da sie der Auffassung aller nicht-radikalen Autonomietheoretiker folgt (Dürscheid, 2016).

Gemein ist beiden Ansätzen (Autonomie- und Interdependenzhypothese), dass sie die geschriebene Sprache als autonomen Forschungsgegenstand und nicht als sekundäre Ausdrucksform der gesprochenen Sprachform betrachten und geschriebene und gesprochene Sprache „als method. differenziert zu behandelnde und theoret. elementare Kategorien

der Sprachbeschreibung und –analyse“ ansehen (Glück & Rödel, 2016, S. 288). Die Interdependenzhypothese grenzt sich insofern von der Autonomiehypothese ab, als dass sie davon ausgeht, „daß [sic.] die gesprochene Sprachform stets das Modell für Verschriftung darstellt“ (Glück & Rödel, 2016, S.288). Während Vertreter der Autonomiehypothese den funktionalen Aspekt der Orthografie betonen und diese in einem hohen Maße strukturieren, wird die Orthografie von Vertretern der Dependenzhypothese nur vordergründig sprachwissenschaftlich fundiert. Orthografische Regularitäten werden als eine linear strukturierte Abfolge von Phonem-Graphem-Zuordnungen dargestellt (Noack, 2001).

Zusammenfassung

Der Gegenstandsbereich der Schriftlinguistik wird im Kontext verschiedener Wissenschaftsdisziplinen jeweils anders betrachtet. Die daraus unterschiedlich resultierenden Definitionen erlauben einen Einblick in die teils kontrovers geführten Diskurse und verdeutlichen die Notwendigkeit einer terminologischen Klärung (Dürscheid, 2016). Die Position, die von den Sprachwissenschaftler*innen innerhalb dieser Kontroverse eingenommen wird, ist jeweils von der Perspektive und vom Forschungsziel abhängig. So ist aus historischer Perspektive zwar das Schreiben dem Sprechen nachgeordnet, aus systemischer Perspektive spielt dieses Argument jedoch keine Rolle (Dürscheid, 2016). Eisenberg (2020a, S.313) beurteilt die Debatte um Autonomie und Dependenz als unfruchtbar, da „immer wieder systematische Fragen mit solchen der historischen Entwicklung, des Erwerbs, der kognitiven Verarbeitung und der gesellschaftlichen Funktion von Schrift und geschriebener Sprache konsequent durcheinandergeworfen“ werden.

2.1.1 Grundeinheiten des Schriftsystems

Alphabetisch verschriftlichte Sprache lässt sich auf der untersten Komplexitätsstufe auf der Phonem- und Graphemebene beschreiben. Die Phonologie und die Graphematik als zwei Teildisziplinen der Sprachwissenschaft untersuchen die verschriftlichte Sprache auf diesen Ebenen (Dürscheid, 2016). Phoneme und Grapheme haben im Gegensatz zu sprachlichen Einheiten, die auf der morphologischen, lexikalischen und syntaktischen Ebene eine bedeutungstragende Funktion haben, nur eine bedeutungsunterscheidende Funktion (Dürscheid, 2016). Während sich die Phonologie mit den Grundeinheiten des Lautsystems beschäftigt, umfasst der Untersuchungsgegenstand der Graphematik die Grundeinheiten des Schriftsystems. Beide Disziplinen untersuchen jeweils die Regeln zur Verknüpfung ihrer Bezugssysteme (Dürscheid, 2016). In der Forschung gibt es keinen Konsens über eine einheitliche Definition des Begriffs „Graphem“. Die unterschiedlichen Definitionen spiegeln vielmehr die Forschungspositionen wider. Über die wesentlichen Gemeinsamkeiten zwischen Graphemen und Phonemen herrscht jedoch Einigkeit. Ein Phonem ist die kleinste segmentale Einheit des Lautsystems, ein Graphem entsprechend die kleinste segmentale Einheit des Schriftsystems (Eisenberg, 2020a). Aus der Sichtweise der Dependenzhypothese wird das Graphem als schriftliche Repräsentation des Phonems und aus der Sichtweise

der Autonomiehypothese als „die kleinste bedeutungsunterscheidende Einheit des Schriftsystems einer Sprache“ ohne Rückbezug auf das Lautsystem beschrieben (Günther, 1988, S.77).

Zur Beschreibung des Schriftsystems auf diesen Ebenen gibt es in der Sprachwissenschaft unterschiedliche Positionen. Besonders populär und im aktuellen Grammatik-Duden niedergelegt sind die Ausführungen von Eisenberg (Eisenberg, 2016). Maas (2010) hingegen formuliert in einigen Bereichen ein konträres Konzept zur Beschreibung des Schriftsystems.

2.2 Graphematische Forschung nach Eisenberg

Im Folgenden wird der Schwerpunkt auf die Schriftsystematik nach Eisenberg (2020a) gelegt. Die graphematische Forschung bietet einen neuen Blick auf die Rechtschreibung und eröffnet neue Möglichkeiten für die Didaktik, indem sie das Konstrukt Rechtschreibung systematisiert und in einen Kern- und einen Peripheriebereich aufteilt. Sie umfasst nicht nur eine Zusammenstellung von orthografischen Regeln auf den Grundlagen der amtlichen Rechtschreibung, sondern auch eine Systematik der Orthografie, die eine Einsicht in die Schriftstruktur der Wörter ermöglicht und die Regularitäten der Wortschreibung erklärbar macht (Eisenberg, 2020a). Die Graphematik wendet sich vom vorherrschenden Konzept der lautgetreuen Schreibung zum Schriftspracherwerb ab, in dem das Geschriebene als Abbild der gesprochenen Sprache gilt. Der Rechtschreiberwerbsprozess dient nicht nur zur Erfüllung gesellschaftlicher Normen. Vielmehr steht das hohe Lernpotenzial dieses Prozesses für die mündliche und schriftliche Sprachkompetenz im Fokus (Blatt, 2010; Eisenberg, 2020a; Eisenberg & Fuhrhop, 2007).

Eisenbergs Graphematik unterscheidet zwischen einem phonographischen, silbischen und morphologischen Prinzip sowie einem Prinzip der Wortbildung. Das Konstrukt der Rechtschreibung wird systematisiert und in einen Kern- und Peripheriebereich aufteilt. Der Kernbereich wird durch das:

- phonographische
- silbische
- morphologische
- und wortübergreifende Prinzip

repräsentiert und stellt ca. 95 Prozent der regulären Schreibungen dar (Eisenberg & Fuhrhop, 2007). Schreibungen, die nicht regelbasiert hergeleitet werden können, wie z.B. das Dehnungs-h oder Fremdwortschreibungen, sind im Peripheriebereich zusammengefasst (Blatt & Prosch, 2016, S. 115). Diese systematische Strukturierung der Orthografie gibt nicht nur die orthografisch korrekte Schreibweise vor, sondern ermöglicht es den Schüler*innen die Sprache und dessen Verschriftlichung verstehbar und nachvollziehbar zu machen (Eisenberg, 2016, S.65):

> „Eine Darstellung dieser Art stellt nicht nur fest, wie geschrieben wird, sondern sie beantwortet auch die Frage nach dem Warum. Sie zeigt, welche allgemeinen Prinzipien der Wortschreibung des Deutschen zugrunde liegen. Der Schreiber kann die Orthografie seiner Sprache nicht nur beherrschen, er kann sie auch verstehen. So wird auch einsichtig, dass die Behandlung der Schriftstruktur sprachlicher Einheiten Teil einer Grammatik des Deutschen sein muss“.
>
> „Weiterhin zeigen die graphematischen Forschungsbefunde auf, dass der zentrale Sinn der Rechtschreibung darin liegt, das Lesen zu erleichtern, ein Aspekt, der seit der Normierung der deutschen Rechtschreibung aus dem Blick geraten ist“ (Blatt, 2010, S.101).

2.2.1 Phonographisches Prinzip

So wie in vielen anderen Sprachen basiert die Schrift des Deutschen auf dem lateinischen Alphabet, das aus insgesamt 26 Buchstaben besteht. Die Grundeinheiten der deutschen Schrift sind die Grapheme. Unter einem Graphem versteht man die kleinste bedeutungsunterscheidende segmentale Einheit des Schriftsystems, unter einem Phonem entsprechend die kleinste segmentale Einheit des Lautsystems. Mehrgrapheme sind Buchstabenverbindungen als kleinste, systematisch unterteilbare Einheiten (Eisenberg, 2016).

Das System des Deutschen besteht aus zwanzig Konsonantenphonemen und sechzehn Vokalphonemen, wobei die Anzahl der Vokalphoneme (insgesamt zwanzig) die der Vokalgrapheme (insgesamt neun) übersteigt (Eisenberg, 2020a). Vokale können gespannt oder ungespannt bzw. lang und kurz ausgesprochen werden. Die Vokalphoneme lassen sich in acht gespannte und sieben betonbare Vokale und in ein unbetonbares (Schwa) untergliedern (Eisenberg, 2020a). Das Phonemsystem dient als lautliche Bezugsgröße für Graphem-Phonem-Korrespondenzen, wobei diese als einfache morphologische Einheit verstanden wird und eine sog. Explizitlautung voraussetzt (Eisenberg, 2020a). Inwiefern ein Zusammenhang zwischen der Struktur des graphematischen mit der des phonologischen Wortes besteht, wird durch sogenannte Graphem-Phonem-Korrespondenzen bestimmt. Der Fokus liegt dabei auf dem Gesprochenen zum Geschriebenen und weniger auf dem Geschriebenen zum Gesprochenen (Eisenberg, 2020a).

Die Grundlage des phonographischen Prinzips bilden Phonem-Graphem-Korrespondenzen, woraus sich häufig die orthographisch korrekte Schreibung ableiten lässt, sofern die sogenannte Explizitlautung als phonologische Bezugsgröße dient. Die Phonemfolge kann Segment für Segment auf die Graphemfolge abgebildet werden (Eisenberg, 2020a). Eisenberg (2020a) formuliert Regeln für Graphem-Phonem-Korrespondenzen (GPK), die angeben, inwiefern graphematische Segmente durch Phonemfolgen repräsentiert werden können. Dabei werden die Graphem-Phonem Zuordnungen als Korrespondenzen und nicht als «Graphemgenerierungsregeln» verstanden, die eine Ableitung von Graphemen aus Phonemen regeln, da die GPK nicht immer zur orthografisch korrekten Schreibungen führen. Eisenberg bezeichnet das Schriftsystem des Deutschen deshalb auch als ein Mischsystem, weil es zudem silbische und logographische Züge aufweist. So spielt die Struktur größerer

sprachlicher Einheiten wie die Silbe, das Morphem und die Wortform bei der orthografisch korrekten Verschriftlichung eines Wortes eine wichtige Rolle (Eisenberg, 2020a).

2.2.2 Silbisches Prinzip

Die Schreibsilbe als Basis des silbischen Prinzips in Abgrenzung zur gesprochenen Silbe bildet den Schwerpunkt der Graphematik Eisenbergs. Diese nimmt in der geschriebenen Form eine andere segmentale Gestalt an als im Gesprochenen, ist im Gegensatz zur Sprechsilbe stärker regularisiert und hat eine größere Formkonstanz, um die Silbenlänge zu vereinheitlichen (Eisenberg, 2020a; Eisenberg, 2016). Die starke silbische Ausrichtung des Deutschen erleichtert das Lesen, da immer identische Ausgleichsmittel in Form von Buchstaben- und Graphemverbindungen auftreten. Die Schüler*innen lernen „bald, solche festen Muster zu erkennen und damit die silbenstrukturelle Information zu erschließen“ (Eisenberg, 2016, S.71). Silben, die sich immer vollständig aus Lauten zusammensetzen, bilden die Grundlage jeder Wortform. Die Silbe als sprachliche Einheit ist zwischen dem Lautsegment und der Wortform anzusiedeln (Eisenberg, 2016). Wortformen werden nicht als Laut-, sondern als Silbenfolgen verstanden (Eisenberg, 2016). Es gibt sowohl betonte als auch unbetonte Silben, „sie sind die Träger von Akzenten und damit von entscheidender Bedeutung für den Sprachrhythmus“ (Eisenberg, 2016, S.38). Die Silbenanzahl und –grenzen sind sowohl in der phonologischen als auch graphematischen Wortform identisch. Dies erleichtert den Lernenden die Einsicht in die syllabische Struktur von Wortformen. Eisenberg (2016, S.37-38) schreibt dazu:

> „Die Gliederung einer Wortform in Silben ist dem Sprecher intuitiv zugänglich. Ohne Schwierigkeiten lässt sich angeben, wie viele Silben eine Wortform hat. Kinder verfügen über diese Kenntnis genauso wie Erwachsene. Bevor Kinder schreiben lernen, wissen sie im Allgemeinen nicht, dass Wortformen aus Lautsegmenten aufgebaut sind. Dagegen machen viele Kinderspiele von der Gliederung der lautlichen Formen in Silben Gebrauch (z. B. Abzählreime).“

Die Silbe besteht aus einem Anfangsrand, Endrand und einem Silbenkern, wobei dieser im Deutschen immer aus einem Vokal oder aus einem Diphthong besteht und obligatorisch ist. Vor und nach dem Silbenkern können maximal drei Konsonanten als Silbenanfangsrand oder Silbenendrand stehen. Einfache Anfangs- und Endränder bestehen aus einem Laut, komplexe Anfangs- und Endränder aus Lauthäufungen (Eisenberg, 2016; Fuhrhop, 2020). Von einer offenen Silbe spricht man, wenn der Endrand einer Silbe unbesetzt, von einer geschlossenen, wenn der Endrand besetzt ist. Eine nackte Silbe liegt dann vor, wenn der Anfangsrand leer ist. Ein bedeckter Anfangsrand liegt vor, wenn dieser mit mindestens einem Konsonanten besetzt ist (Fuhrhop, 2020). Im Folgenden werden die silbenstrukturellen Merkmale beschrieben, die zu den genannten Abweichungen in der Graphem-Phonem-Korrespondenz führen. Obwohl der Silbenanfangsrand in fast allen Fällen phonographisch geschrieben wird, existiert mit der Verkürzung des Silbenanfangsrands eine regelhafte Abweichung vom phonographischen Prinzip zugunsten der Silbenstruktur. So wird das Phonem in Wörtern wie z.B. Strich und Splitter konträr

zur Graphem-Phonem-Korrespondenz (GPK) zur Vermeidung einer Überlänge im Schriftbild nicht mit <Schtr> und <Schpl>, sondern mit <Str> und <Spl> verschriftlicht. Im Anfangsrand einer Silbe können maximal drei Konsonantenbuchstaben stehen. Verkürzte Silbenanfangsränder verhindern graphematische Überlängen, erhalten dadurch die Übersichtlichkeit einer Schreibsilbe und erleichtern das Lesen trotz Abweichungen (Eisenberg, 2016).

Im Deutschen besteht eine Interdependenz zwischen dem Silbenaufbau und der Vokalquantität bzw. -qualität. Die Vokallänge eines Wortes kann auf der Grundlage des Aufbaus der Silbe erschlossen werden und es bedarf keiner besonderen Markierung, ob ein Vokal lang oder kurz gelesen wird (Eisenberg, 2016). Eine systematische Ausnahme stellt das <ie> dar. Für betonte Silben des Kernwortschatzes gilt folgendes (Eisenberg, 2016):

- gespannte Vokale sind lang und ungespannte Vokale kurz
- ist der Endrand der Silbe unbesetzt, handelt es sich um eine offene Silbe, der Vokal ist lang wie z.B. in der ersten Silbe des Wortes <Rose>
- ist der Endrand der Silbe mit mindestens zwei Konsonanten besetzt, handelt es sich um eine geschlossene Silbe und der Vokal ist kurz, wie z.B. in dem Wort <List>

Bei betonten Silben mit einfachem Endrand gelten besondere Regeln für lange, gespannte und kurze und ungespannte Vokale (Eisenberg, 2016). Für flektierte Einheiten gilt:

- „Hat ein Einsilber nur ein Graphem im Endrand, so wird der Vokal lang gelesen, z. B. Ton, Flut, schön, groß.
- Hat eine Silbe im Mehrsilber ein Graphem im Endrand, so wird sie kurz gelesen, z. B. Mul-de, Kan-te, Gür-tel, Wol-ke (Ausnahme: Wüs-te)“ (Eisenberg, 2016, S.73).

Die Vokalquantität in Einsilbern kann anhand von Langformbildungen ermittelt werden. Das Silbengelenk stellt eine weitere explizite grafische Markierung der Silbengrenze dar. Liegt ein Silbengelenk vor, so wird dieses „auf zwei graphematische Einheiten zerdehnt“. Von einem Silbengelenk wird nach Eisenberg (2016, S.76) gesprochen, wenn „in einer phonologischen Wortform zwischen einem betonten ungespannten und einem unbetonten Vokal ein einzelner Konsonant“ steht. So wird der Buchstabe <t> in dem Wort <Schlitten> zur ersten und zur zweiten Silbe gezählt. In der deutschen Orthografie ist es nicht möglich, dass ein Segment gleichzeitig zur zweiten Silbe gezählt wird. Der Silbenschnitt liegt dabei zwischen den zwei Konsonanten. Die Verdopplung eines Konsonanten dient folglich nicht zur Markierung der Kürze eines Vokals, sondern zur Markierung von Silbengelenken. Silbengelenke treten nur nach Kurzvokalen auf, jedoch stellt dies keine Erklärung für das Auftreten des Doppelkonsonanten dar. Eisenberg (2016) führt zur Verdeutlichung die Wörter <in>, <von>, <um>, <ab> an und zeigt, dass in diesen Wörtern zwar Konsonanten nach Kurzvokalen auftreten, jedoch keine Silbengelenke vorhanden sind und im Geschriebenen dazu keine Gemination stattfindet.

Das silbeninitiale <h> stellt eine explizite grafische Markierung der Silbengrenze dar und tritt immer dann auf, wenn zwei Silbenkerne aufeinandertreffen. Dies ist häufig der Fall

bei betonten offenen Silben. Das <h> wird an den Anfang der zweiten Silbe vorangestellt. Es markiert nicht nur die Silbengrenze, sondern unterstützt die visuelle Prägnanz der graphematischen Wortform. Dieser Funktion kommt beim Auftreten von Vokalbuchstabenhäufungen besondere Bedeutung zu. So würden beispielsweise die Wörter <ziehe>, <fliehe> ohne das silbenininitale <h> als <ziee> und fliee> verschriftlicht (Eisenberg, 2016). Ebenso wie das Dehnungs-h entspricht das silbeninitiale <h> keinem Phonem und ist stumm. Es wird nicht gesetzt, wenn bereits zwei Vokalbuchstaben in einem Wort vorkommen, wie z.B. bei den Wörtern (See-Seen, Knie–Knie, freu-en, trau-en) (Eisenberg, 2016, S.76). Beim Diphthong <ei> kommt es zu einer Abweichung. Es gibt Fälle, in denen das <h> gesetzt wird, obwohl zwei Vokalbuchstaben vorhanden sind (z.B. Reihe, Reiher, Weihe). In Fällen, in denen es nicht gesetzt wird (z.B. Schreie) gilt: ist die Grundform einsilbig (z.B. Schrei-Schreien, Blei-verbleien), wird das silbeninitiale <h> nicht verschriftlicht (Eisenberg, 2016, S.76).

2.2.3 Morphologisches Prinzip

Nicht nur silbenstrukturelle, sondern auch morphematisch bedingte Eigenschaften führen zur Abweichung der Graphem-Phonem-Korrespondenzen. Das deutsche Schriftsystem wird aufgrund dessen auch als tiefes System bezeichnet (Dürscheid, 2016). Der morphologische Teil des Schriftsystems basiert auf der phonologischen Schreibung einfacher morphologischer Einheiten. Morphologisch komplexe, phonologische Wortformen unterliegen häufig spezifischen Assimilations- und Reduktionsprozessen, die nur im Gesprochenen auftreten. Unter einem Morphem wird eine bedeutungstragende Einheit verstanden, die die Grundlage für Wortformen darstellt. Es gibt sowohl Wortformen, die aus genau einem Morphem bestehen, wie z.B. bei, schnell, Hand als auch welche, die aus mehreren Morphemen bestehen (z.B. da#bei, schnell#er#es, hand#lich) (Eisenberg, 2016). Die Struktur der sprachlichen Einheit des Morphems spielt eine wichtige Rolle für die orthografisch korrekte Schreibweise. Während die Lautgestalt eines Morphems durch Flexionen und Ableitungen einer gewissen lautlichen Veränderung unterliegt, bleibt die orthografische Schreibung meist relativ gleich (Eisenberg, 2016). Die Morphemkonstanz beschreibt, dass sich das Geschriebene klar vom Gesprochenen unterscheidet und die Segmentfolge der geschriebenen morphologischen Einheit in fast allen Fällen konstant bleibt (Eisenberg, 2016) und sowohl für das Lesen als auch das Schreiben eine Hilfe darstellt. Die graphematische Form und die Bedeutung des Morphems lassen sich direkt aufeinander beziehen. Das Morphem kann so durch seine Konstanz leichter identifiziert werden (Dürscheid, 2016, Nerius, 2007). Nerius (2007, S.149) weist wie folgt auf diesen Sachverhalt hin:

> „Der versierte Schreibende greift nämlich beim Produzieren von graphischen Morphemformen (auch in neuen Wortbildungsprodukten) im Allgemeinen auf bereits gespeicherte Muster, auf die graphischen Erinnerungsbilder zurück“.

Im Deutschen lassen sich prosodische und morphologische determinierte Explizitformen identifizieren.

> „Prosodisch determinierte Explizitformen stellen die Basis für eine Konstantschreibung morphologischer Einheiten und insbesondere von Stammformen bereit, deren Variation im Gesprochenen rein phonologischer Natur ist“ (Eisenberg, 2020a, S.343).

Unter prosodischen Explizitformen werden Schreibungen verstanden, „die auf einen Trochäus oder Daktylus enden, wobei die Ultima bzw. Pänultima nicht ein konsonantisch anlautendes Suffix/Pseudosuffix (Endung) enthält“ (Eisenberg, 2020a, S.339). Ein prominentes Beispiel hierfür stellt die Auslautverhärtung dar. So würde das Wort <Kind> phonographisch korrekt als <Kint> verschriftlicht. Jedoch erklärt sich die Schreibung der Auslautverhärtung durch den Rückbezug auf die zweisilbige Form des Genitivs <Kindes>, woraus sich die Stammschreibung <Kind> ergibt. Die Explizitform kann auch aus einer Pluralform wie <Kinder> oder <Kindern> resultieren, vorausgesetzt, die geforderte prosodische Struktur ist vorhanden (Eisenberg, 2020a). Eine solche Konstantschreibung im Kernwortschatz gilt „für alle Formen innerhalb der Flexionsparadigmen von Substantiven, Adjektiven und Verben“ und für Stammformen in abgeleiteten Wörtern (Eisenberg, 2020a, S.339):

> „Ist eine Gelenkschreibung, ein silbeninitiales <h>, ein Dehnungs-h oder ein Doppelvokalgraphem Bestandteil einer phonologischen Schreibung, dann bleiben sie in fast allen morphologischen verwandten Formen erhalten.“

Bei der morphologisch determinierten Explizitform spielen lautliche Unterschiede zwischen Stammformen eine Rolle: „Die Grapheme ⟨ä⟩, ⟨ö⟩, ⟨ü⟩ zeigen durch ihre Form an, dass der Umlaut morphologisch bestimmt sein kann. Die Zusammengehörigkeit von Formen wie *flach – Fläche – flächig; Not – Nöte – nötig; Fluss – Flüsse – flüssig* wird durch die Umlautschreibung hervorgehoben. Dasselbe gilt für den Diphthong ⟨äu⟩ wie in *Haus– Häuser, Schaum– schäumen*“ (Eisenberg, 2016, S.80).

Ferner gibt es in der deutschen Orthografie weitere Schärfungs- und Dehnungsgrafien (Dehnungs-h und die Verdoppelung von Vokalgraphemen), die die Vokallänge bei bestimmten Typen von Schreibsilben anzeigen (Eisenberg, 2020a). Sowohl Doppelvokale als auch das Dehnungs-h stehen in der offenen betonten Silbe als Längenmarkierung, obwohl sie zur Dehnung des Vokals nicht benötigt werden. Dieser wird auch ohne das h lang ausgesprochen. Schärfungs- und Dehnungsgraphien zählen zum Randbereich der deutschen Rechtschreibung. Sowohl in einsilbigen als auch in zweisilbigen Wörtern tritt das Dehnungs-h auf, wenn einem Vokalgraphem ein Sonorantengraphem (l, m, n, r) folgt.Es dient eher als Lesehilfe als zur Anzeige einer Dehnung, da der Vokal in der betonten Silbe ohnehin lang gelesen wird. Das Dehnungs-h als Längenmarkierung ist ein stummes <h>, da diesem Graphem kein Phonem entspricht.

Nach einem Sonoranten können noch weitere Konsonanten folgen. Sie treten besonders häufig am Anfang komplexer Endränder auf wie z.B. in den Wörtern <Welt>, <Furcht>, <Hirn> etc.. Das Auftreten eines Sonoranten nach dem Silbenkern weist auf einen komplexen Endrand hin und auf die Länge des vorausgegangenen Vokals. Hier zeigt sich die Bedeutung des Dehnungs-h als Lesehilfe. Ferner trägt es (Dehnungs-h) in morphologisch einfachen Formen mit einem einfachen Endrand zu einem Ausgleich der Schreibsilbe bei,

da es den Endrand optisch verlängert. Es lassen sich Annäherungsregeln definieren: Steht nach einem Vokalgraphem ein Sonorantengraphem (r,l,m,n), wird in ca. 50% der Fälle das Dehnungs-h geschrieben. Zwar stellt das Auftreten des Dehnungs-h vor einem Sonorantgraphem (r, l, m, n) eine notwendige Bedingung dar, jedoch keine hinreichende, da es in vielen Fällen nicht auftritt, wo es auftreten könnte. Eisenberg (2016, S.74) formuliert in Abhängigkeit zur Silbenlänge Tendenzen zum Auftreten des Dehnungs-h:

- „Bei Silben mit komplexem Anfangsrand ist das Dehnungs-h eher unwahrscheinlich, vgl. *Strom, schwer, Schwan, Schnur, schwül.*
- Bei Silben mit einfachem Anfangsrand ist das Dehnungs-h dagegen eher wahrscheinlich, vgl. *Hahn, hohl, kahl, Ruhm, Wehr, kühn.* Auch hier besteht also die Tendenz zum Ausgleich der optischen Silbenlänge."

Die Verdopplung von Vokalgraphemen dient wie das Dehnungs-h als Lesehilfe und führt zu einem Ausgleich der Schreibsilbe. Anhand der Verdopplung eines Vokalgraphems kann nicht auf die Vokallänge geschlossen werden, da dieses Phänomen nur in jenen Fällen auftritt, in denen ein Vokalgraphem ohnehin lang gelesen wird. Eine Vokalgraphemverdopplung kann in den folgenden drei Kontexten auftreten:

- „⟨ee⟩ steht in offener Silbe wie in *Schnee, Tee, See, Fee*
- ⟨aa⟩, ⟨ee⟩, ⟨oo⟩ treten auf vor ⟨r⟩ und ⟨l⟩ wie in *Aal, Saal, Haar, Paar, scheel, Heer, Meer, Teer, leer, Moor*
- ⟨aa⟩, ⟨ee⟩, ⟨oo⟩ stehen vor ⟨t⟩ wie in *Staat, Saat, Maat, Beet, Boot*, auch wenige Einheiten mit ⟨s⟩ gibt es *Aas, Moos*, dazu den Einzelfall *Waage*. Neben den Sonoranten sind [t] und [s] diejenigen Laute, die am häufigsten in komplexen Endrändern und damit nach kurzem, ungespanntem Vokal vorkommen" (Eisenberg, 2016, S.74).

Während der erste Kontext am häufigsten auftritt, kommen die anderen eher selten vor. Eisenberg (2016) vermutet, dass die Verdopplung der Vokale <i> und <u> aufgrund ihrer Form in der deutschen Orthografie nicht vorkommen, da dies zu einer Irritation der Augen führen könnte.

2.2.4 Prinzip der Wortbildung und Wortformbildung

Ein weiterer Teil der Morphologie sind Wortbildungen und Wortformbildungen. Durch Wortbildungen wird der Wortschatz auf der Grundlage bereits vorhandener Sprachelemente erweitert. Eisenberg (2016) unterscheidet zwischen den vier Regularitäten Komposition, Präfigierung, Suffigierung und Konversion als Grundlage dafür. Bei Kompositionen treffen zwei Morpheme aufeinander wie z.B. in den Wörtern Edelmut, Fensterrahmen, Baukran, Geisteswissenschaft. Zwischen zwei freien Morphemen tritt in 27 Prozent aller Substantivkomposita und 30 Prozent aller Adjektivkomposita ein Fugenelement (<s>,<n>, <ns>, <e>, <er>, <en>, <es> und <ens>) auf (Eisenberg, 2016). Wird ein Wortbildungsaffix dem Stamm vorangestellt, liegt eine Präfigierung vor, folgt ein Wortbildungsaffix einem Wortstamm, liegt eine Suffigierung vor z.B. <mutig>, <freundlich>, <Ehrung>,

<Sicherheit>. Bei einer Konversion bleibt der Wortstamm unverändert und wird in eine andere Kategorie umgesetzt. Im Deutschen zählen die Substantivierung des Infinitivs <das Lesen>, <das Wandern>, <das Maßhalten> und die Substantivierung des Adjektivs <der Alte>, <die Abgeordnete>, <das Unvergessene> zu den wichtigsten Arten der Konversionen (Eisenberg, 2016).

Ein weiterer zentraler Bestandteil der deutschen Grammatik sind Flexionen. Bei Wortformenbildungen spielen Deklinationen, Konjugationen und Komparationen eine Rolle, die die Veränderung eines Wortes bezüglich der grammatischen Funktion anzeigen. Zu den deklinierten Wortarten zählen die Substantive, Adjektive, Pronomina und Artikel. Werden Verben flektiert, so spricht man von Konjugationen (Eisenberg, 2016). Mittels Komparationen lassen sich Adjektive steigern.

2.2.5 Sprachsystematische Rechtschreibdidaktik

Um die Potenziale der graphematischen Forschung für eine Rechtschreibdidaktik nutzen zu können, bedarf es großer Investitionen (Eisenberg & Fuhrhop, 2007).

Hinney (1997, 2004) kommt mit ihrer „Neubestimmung von Lerninhalten für den Rechtschreibunterricht“ der Verdienst zu, die Graphematik nach Eisenberg in die Rechtschreibdidaktik integriert zu haben. In ihrem didaktischen Ansatz sind Lerninhalte und -ziele in Anlehnung an Eisenbergs phonographische, silbische und morphologische Prinzipien formuliert, die den Ausgang für Schriftlernende bilden. Einen besonderen Stellenwert erhält das silbische Prinzip. Erst danach soll der Peripheriebereich behandelt werden. Ein lautorientierter Zugang allein ist für die Herleitung von Wortschreibungen ungeeignet, da es viele Abweichungen durch Sonderregelungen und Ausnahmeschreibungen gibt (Hinney, 2015; Hinney & Menzel, 1998). Zudem können die Regeln nach dem Phonem-Graphem-Korrespondenz-Prinzip eher von bereits Schriftkundigen verstanden werden (Hinney, 2015).

Viele Kinder haben Schwierigkeiten, die Lang- und Kurzvokale zu hören, die jedoch in den von der Mündlichkeit ausgehenden didaktischen Modellen im Fokus des Lernens stehen. Hinney und Menzel (1998, S.272) führen dazu aus:

> „Der Einsatz der bloßen Mündlichkeit zur Lösung rechtschreiblicher Probleme ist theoretisch und praktisch einfach unbrauchbar. Die Annahme, lautgetreue Wortschreibungen würden ohne eine ungefähre Einsicht in den Aufbau von Wörtern überhaupt möglich sein, kann nur von dem vertreten werden, der die Wortschreibung kennt.“

Hinney (1997) entwickelt ein Konstruktionsprinzip der Wortschreibung in zwei Schritten. Es stellt den Kindern sprachanalytische Untersuchungsmethoden zur Verfügung, mit denen sie den Aufbau von Schreibungen selbst entdecken können (Hinney, 2004; Hinney, 2014). Grundlage sind die Regularitäten der Schreibsilbe im prototypischen Zweisilber (zweisilbige Langform), die fast alle Merkmale der Wortschreibung (z.B. Vokallänge und -kürze, Silbengelenk, silbeninitiales h, ⟨ie⟩-Schreibung) erklären können. Der erste Schritt

umfasst die Beschäftigung mit phonologischen Gesetzmäßigkeiten (phonographisches und silbisches Schreiben) über eine Silbenprobe des prototypischen Zweisilbers (Hinney, 2004). Schüler*innen analysieren die geschriebene und gesprochene Form, um z. B. den Unterschied in der Vokallänge bzw. -kürze der Phoneme /U/ und /u:/ in Mutter und Musik durch den Silbenschnitt bei Mutter und Musik herleiten zu können. Der zweite Schritt ist den Gesetzmäßigkeiten des Wortaufbaus und des morphologischen Prinzips gewidmet. Es geht darum, dass vererbte silbenstrukturelle Informationen deutlich werden (Hinney, 2004). Sie verhelfen Schriftlernenden herauszufinden, dass z.B. das flektierte Wort <renn> mit ⟨nn⟩ geschrieben wird, wegen <ren-nen>, oder der Einsilber Kind mit ⟨d⟩ wegen Kinder (Naujokat, 2015).

2.3 Zusammenfassung

Die graphematische Forschung bietet wichtige Erkenntnisse und Einsichten, die beim Erlernen der Rechtschreibung hilfreich sein können. Sie wendet sich vom vorherrschenden Konzept der lautgetreuen Schreibung zum Schriftspracherwerb ab, in dem das Geschriebene als Abbild der gesprochenen Sprache gilt. Der Rechtschreiberwerbsprozess dient nicht nur zur Erfüllung gesellschaftlicher Normen. Vielmehr steht das hohe Lernpotenzial dieses Prozesses für die mündliche und schriftliche Sprachkompetenz im Fokus (Blatt, 2010; Eisenberg, 2020a; Eisenberg & Fuhrhop, 2007). Die Erkenntnisse der graphematischen Forschung können dazu beitragen, Schüler*innen das Verständnis für das Schriftsystem und dessen Struktur zu vermitteln. Sie umfasst nicht nur eine Zusammenstellung von orthografischen Regeln auf den Grundlagen der amtlichen Rechtschreibung, sondern auch eine Systematik der Orthografie, die eine Einsicht in die Schriftstruktur der Wörter ermöglicht und die Regularitäten der Wortschreibung erklärbar macht (Eisenberg, 2020a).

3 Schriftspracherwerb und Rechtschreibkompetenz

Dieses Kapitel erschließt die Potenziale einer sprachsystematischen Modellierung von Rechtschreibkompetenz als Grundlage für die Konstruktion eines Instruments zur Lernverlaufsdiagnostik und deren Implikationen für den Schriftspracherwerbsunterricht. Darüber hinaus thematisiert das Kapitel die Bildungsstandards für das Fach Deutsch in der Primarstufe und die Diskussion um den Kompetenzbegriff in der empirischen Bildungsforschung sowie die Anforderungen, Rechtschreibkompetenz zu modellieren und zu messen, da diese Aspekte bei der Konstruktion eines Instruments zur kompetenzorientierten Lernverlaufsdiagnostik eine Rolle spielen.

3.1 Entwicklungsmodelle des Schriftspracherwerbs

Der Schriftspracherwerb stellt einen komplexen Prozess der Denkentwicklung dar, bei dem die Schüler*innen das Schriftsystem entsprechend ihrer kognitiven Entwicklung verstehen und rekonstruieren müssen (Weinhold et al., 2020). Durch diesen Prozess gewinnen die Schüler*innen ein tieferes Verständnis des Zusammenhangs zwischen gesprochener und geschriebener Sprache, ihre Sprachkompetenz und ihre kognitiven Strukturen entwickeln sich weiter (Weinhold et al., 2020). Das Thema Schriftspracherwerb steht seit etwa 40 Jahren im Fokus der sprachdidaktischen Forschung. Während im englischsprachigen Raum seit den 1980er Jahren Studien zur Überprüfung von Modellen zum Schriftspracherwerb durchgeführt werden, blickt die Forschung im deutschsprachigen Raum auf eine vergleichsweise kurze Tradition zurück. Einen ausführlichen Überblick sowie eine kritische Bestandsaufnahme und Analyse bestehender Modelle zum Schriftspracherwerb findet sich bei (Becker, 2008). Im Kontext der vorliegenden Arbeit sollen nur wichtige Aspekte der Modelle vorgestellt werden. Ein erstes Modell, das Schreiben als einen Entwicklungsprozess versteht, stammt von Eichler (1976). Er spricht von einem „inneren Regelbildungsprozess“, der die vielen impliziten Aktivitäten beschreiben soll, die damit verbunden sind. Das sechsstufige Modell von Frith (1986) dient bis heute als Vorlage für weitere Entwicklungsmodelle (Fay, 2013). Diese unterscheiden sich hinsichtlich der Integration bzw. Segregation des Lesenlernens und Schreibenlernens (z.B. Brügelmann & Brinkmann, 1994; Dehn, 1985; May et al., 2002; Scheerer-Neumann, 2008; Spitta, 1988; Thomé, 2003; Valtin, 2000). Den Modellen liegt die Annahme zugrunde, dass der Schriftspracherwerb eine strikte Abfolge von Stufen und Strategien ist, die vom Leichten (Phoneme) zum Schweren (orthografische Regeln) führen. Meist wird zwischen einer logographischen, einer alphabetischen und einer

orthografischen Entwicklungsphase unterschieden. Die logographische Phase beschreibt die Nachahmung des Schreibens in Form von Kritzelschrift oder wahllosem Schreiben von Buchstaben ohne Lautbezug. Die alphabetische Phase beschreibt die Fähigkeit, einzelnen Sprachlauten Buchstaben zuordnen zu können. In der orthografischen Phase sind die Lernenden in der Lage, Rechtschreibregeln anzuwenden, die sich nicht auf das phonographische Prinzip beziehen (Fay, 2013). Neuere Ansätze (z.B. Brügelmann & Brinkmann, 1994; Scheerer-Neumann, 2008; Valtin, 2000) beziehen sich auf Erkenntnisse der kognitiven Entwicklungspsychologie, nach denen der Eigenaktivität der Lernenden beim Aufbau der orthografischen Domäne eine hohe Bedeutung zukommt.

Die jeweils unterschiedlichen Modelle sind immer auch Ausdruck „einer ganz bestimmten Perspektive auf den Lerngegenstand Orthografie“, teilweise sehr eingeschränkt und stimmen inhaltlich nur bedingt überein (Budde et al., 2012, S.122). Die aus den Modellen abgeleiteten Konsequenzen und Implikationen sind „in einem hohen Maß von den sprach- und lerntheoretischen Vorannahmen abhängig“(Hinney, 2010, S.50). Durch fachwissenschaftliche Ansätze werden ordnende Muster geschaffen, d.h. „Konstruktionen, die die Vielfalt der historisch gewachsenen Schreibweisen im Nachhinein auf wenige Regelhaftigkeiten zurückführen“ (Brügelmann & Brinkmann, 2013, S.2). Dabei handelt es sich jedoch immer nur um mögliche Modelle, denn keines kann alle Schreibweisen erklären. So konkurrieren unterschiedliche Modelle miteinander, „die jeweils bestimmte Aspekte zu Lasten anderer hervorheben (z.B. Sprechsilbe vs. Morphem [lau-fen vs. lauf-en])“ (Brügelmann & Brinkmann, 2013, S.2). Die Frage, zu welchen Anteilen die Schriftsprache erworben bzw. erlernt wird, ist innerhalb der sprachdidaktischen Forschung nach wie vor ungeklärt (Fay, 2013). Aus einer entwicklungspsychologischen Perspektive auf Rechtschreibung im Sinne von Piaget, die von einem Erwerb mittels innerer Hypothesen- und Regelbildungsprozesse ausgeht, könnten die erwähnten Entwicklungsmodelle als Erklärung für diesen Erwerbsprozess dienen (Fay, 2013). Lernprozesse sind immer auch unterrichtsgeleitet und der Schriftspracherwerb hängt mit der Art und Weise des Rechtschreibunterrichts zusammen(Fay, 2013). Dementsprechend sind Entwicklungsmodelle immer auch als Ergebnis didaktischer Bemühungen zu verstehen, „da die Didaktik damit etwas erklärt, was sie methodisch selbst verursacht hat“ (Bredel et al., 2017, S.96). Die Ausrichtung vieler Kompetenzmodelle (z.B. May et al., 2002) an den amtlichen Regelungen der deutschen Rechtschreibung (vgl. Kap. 2.1), die die deutsche Standardaussprache bei der 1:1-Zuordnung von Graphemen zu Phonemen zugrunde legt, ist in vielfacher Weise problematisch (Hinney, 2010). Die amtlichen Regelungen der deutschen Rechtschreibung basieren auf der Sichtweise der Dependenzhypothese. Wie bereits gezeigt (vgl. Kap. 2), können sie der Struktur der Orthografie nur teilweise gerecht werden. Sie folgen der über Jahrhunderte tradierten Maxime „Schreib wie du sprichst“. Das amtliche Regelwerk der deutschen Rechtschreibung stellt zwar eine hilfreiche Funktion für Schriftkundige dar, kann aber bei Schreiblernenden, die im Rahmen des Schriftspracherwerbs erst ein metakognitives Sprachbewusstsein entwickeln müssen, zu großen Unsicherheiten und Fehlschreibungen führen (Hinney, 2010; Noack, 2001). So können z. B. die Wörter <haben> oder <aber> nach dieser Maxime auch als *habn* und *aba* geschrieben werden (Hinney, 2015). Eine weitere Kritik ist, dass Entwicklungsmodelle, die sich an den amtlichen Regelungen der deutschen Rechtschreibung ausrichten, Entwicklungsphasen beschreiben, die Kom-

petenzdimensionen umfassen, die analytisch zu trennen wären (Fay, 2013). Am Modell von Frith (1986) wird zum einen die Übertragung aus dem Englischen ins Deutsche kritisiert, zum anderen, dass die postulierten Stufen den individuellen Lernverläufen nicht gerecht werden können (Bulut, 2019; Klicpera et al., 2020; Weinhold et al., 2020). Der Erwerb von Rechtschreibkompetenz ist ein komplexer Prozess, der mit individuellen Entwicklungsverläufen bei den Schüler*innen einhergeht (Bulut, 2018; Hanke & Schwippert, 2005; Weinhold et al., 2020). Ergebnisse aus Studien zur längsschnittlichen Entwicklung des Schriftspracherwerbs, wie z.B. zu Entwicklungsmustern von (schwachen) Rechtschreibleistungen und individuellen Schriftlösungen (EnTleS), bestätigen immer wieder, dass die Entwicklungsschritte hin zu einem rechtschreibkompetenten Schreiben sehr heterogen und die individuellen Schreiblösungen sehr unterschiedlich sind (Weinhold et al., 2020). Viele Schüler*innen haben häufig noch in der dritten Klasse Schwierigkeiten bei der lautorientierten Schreibung und bei der orthografisch korrekten Verschriftung (Naumann, 2015). Ihre Schreiblösungen sind nicht kontinuierlich, basieren eher nicht auf einem transferfähigem und systematischem Wissen und können als Ausdruck „des immer wieder Neu-Anfangens“ angesehen werden (Weinhold et al., 2020, S.27). Zwar lassen sich gemeinsame Entwicklungstendenzen bei den Schüler*innen feststellen, die Schlussfolgerung, dass der Schriftspracherwerb linear verläuft, wie in den Modellen postuliert, ist jedoch nicht korrekt:

> „Denn die starken Fluktuationen eines Lernenden gerade zu Beginn des Schriftspracherwerbs (Brinkmann, 2003) zeigen, dass die Rechtschreibung weniger eine Abfolge unterschiedlicher Stufen, sondern vielmehr ein kontinuierlicher verlaufender, interaktiver Prozess ist, bei dem sich der Schwerpunkt der dominierenden Strategie individuell verschiebt (Brinkmann, 1997)“ (Bulut, 2019, S.41-42).

Auch die Annahme eines wechselseitigen Lese- und Schreibprozesses in Entwicklungsmodellen wie dem von Frith (1986) wird kritisiert. Forschungsergebnisse deuten darauf hin, dass es keine enge Verzahnung zwischen Lesen- und Schreibenlernen gibt (Becker, 2008). Dagegen spricht auch, dass es zwar Lesen ohne Schreiben, aber kein Schreiben ohne Lesen gibt (Becker, 2008). Im Vergleich zum Lesenlernen erfordert das Schreibenlernen mehr explizite und kontrollierte Prozesse. Lesen hingegen kann unabhängig vom Schreiben erworben werden (Becker, 2008). Zudem führt das Erreichen einer höheren Lesekompetenz nicht zu einer höheren Rechtschreibkompetenz.

3.2 Rechtschreibkompetenzmodelle

Für die Definition und Modellierung von Rechtschreibkompetenz werden verschiedene Erwerbsmodelle zum Schriftspracherwerb als Ausgangspunkt gewählt. Unterschieden wird zwischen normorientierten, statischen und strukturorientierten prozessualen Kompetenzmodellen (Jagemann & Weinhold, 2018).

Im Rechtschreibunterricht wird häufig auf Kompetenzmodelle zurückgegriffen, die sich an den amtlichen Regeln der deutschen Rechtschreibung orientieren und eine 1:1-Zuordnung

von Lauten und Buchstaben nach dem Motto „Schreibe, wie du es sprichst“ suggerieren und ein metakognitives Sprachbewusstsein voraussetzen (Hinney, 2010, S. 58; Weinhold et al., 2020). Dieses Vorgehen ist in zweierlei Hinsicht problematisch. Zum einen kann der in Modellen zum Schriftspracherwerbsprozess häufig als stufenförmig angenommene Entwicklungsprozess vom einfachen, lautorientierten Schreiben zum komplexeren, orthographischen Schreiben nur als „eine starke Abstraktion des eigentlichen Lernprozesses“ angesehen werden (Bulut, 2019, S. 39). Die Annahme eines solchen linearen und stufenförmigen Erwerbs gilt im Fachdiskurs als widerlegt (Bulut, 2019; Klicpera et al., 2020; Weinhold et al., 2020). Vielmehr wird inzwischen von einem sich allmählich entwickelnden Parallelerwerb ausgegangen (Böhme et al., 2017). Diese Erkenntnisse stehen im Gegensatz zu den häufig postulierten Entwicklungsstufen in Rechtschreibkompetenzmodellen, wie z.B. dem Modell von May (2013), das von der zentralen Annahme ausgeht, dass es grundlegende Strategien zur Schreibung von Wörtern und Sätzen gibt, die im Laufe des Rechtschreiblernens schrittweise erworben werden müssen (May, 2013). Zum anderen ist die Sprachbewusstheit zentral für die Entwicklung der Rechtschreibkompetenz. Sowohl Kinder mit Deutsch als Zweitsprache als auch Kinder mit Problemen im Schriftspracherwerb weisen „eine rudimentär ausgebildete Sprachbewusstheit und eine ideolektgeprägte Ausspracheform“ auf (Hinney, 2010, S.58). Dies stellt eine erschwerte Ausgangslage für den Schriftspracherwerb dar. Bei diesen Schüler*innen führt diese Vorgehensweise zu unleserlichen Schreibungen. Sie können daher keine adäquate Vorstellung von der Struktur der Schrift aufbauen und verlieren so bereits zu Beginn des Schriftspracherwerbs den Anschluss (Blatt & Pagel, 2009). Eine Orientierung an der Standardlautung setzt ein metalinguistisches Prozesswissen voraus, nämlich „das Wissen um die Segmentierung der Laute und deren Klassifikationen als Phoneme“ das „...Paradigma der bloßen Mündlichkeit, [das eine schreibnützliche- bzw. Pilotsprache voraussetzt] als Basis für Schriftlichkeit ist die Perspektive des Schriftkundigen“ (Hinney, 2010, S.58-59). Auch Günther (2010, S.15) beurteilt, ähnlich wie Hinney (2010) einen Schriftspracherwerbsunterricht, der lediglich auf die „Vermittlung eines anderen Kanals“ abzielt, im Sinne eines knowing how, das die Kinder im Bereich der Lautsprache schon haben, in ein „*knowing that* zu transformieren und dieses dann auf die Schrift anzuwenden“, als keinen adäquaten Ansatz für den Schriftspracherwerb. Denn einfache kognitiv-sprachliche Leistungen sind ohne das Verstehen von Schrift nicht möglich und die eindimensionale Abbildung von Mündlichkeit auf Schriftlichkeit greift zu kurz (Günther, 2010, S.15). Günther (2010) kritisiert die Annahme, dass die Schüler*innen nur das Prinzip der Alphabetschrift verstehen lernen müssten und dass sie bereits über das notwendige Wissen verfügen würden und nur noch die Buchstabenformen auf die ihnen zur Verfügung stehende Phonologie abbilden müssten. Die Zusammenhänge beim Schriftspracherwerb sind jedoch umgekehrt. Erst durch

> „das Verständnis der elementaren Eigenschaften geschriebener Zeichen und Texte [...] entsteht [...] die Möglichkeit, sich ein Bild von der Lautsprache und ihrer Struktur zu machen [...] Zunächst aber fehlt phonologisches Bewußtsein. Erst mit dem Erwerb der Möglichkeit, sich Lautäußerungen als Folge diskreter Teile vorzustellen, kommt es zu Neuentdeckungen, die ihrerseits wieder das Schriftverständnis fördern“ (Günther, 2010, S.15-16).

Die Konsequenzen für einen Unterricht zum Schriftspracherwerb sind, dass an die Vorer-

fahrungen der Kinder angeknüpft werden muss und dass der Unterricht die unterschiedlichen Niveaus berücksichtigen muss, denn „es gibt keine Stunde Null für den Schriftspracherwerb - weder in dem Sinne, daß alle Kinder die gleichen Voraussetzungen hätten, noch in dem Sinne, daß alle Kinder gar nichts wüßten“(Günther, 2010, S.16). Ein Schriftspracherwerbsunterricht, der den Gegenstand Rechtschreiben angemessen modelliert und als systematisch und lernbar vermittelt, ist zentral für einen erfolgreichen Schriftspracherwerb und für die Unterstützung des komplexen Prozesses der Denkentwicklung. Davon profitieren Schüler*innen mit Problemen beim Schriftspracherwerb in besonderem Maße (Weinhold et al., 2020).

Inwiefern sich die unterschiedlichen schrifttheoretischen und -didaktischen Positionen in den Modellen zum Schriftspracherwerb niederschlagen, soll im Folgenden anhand des Rechtschreibkompetenzmodells von May (2013) und dem validierten sprachsystematischen Rechtschreibkompetenzmodell verdeutlicht werden.

3.2.1 Normbasierte Rechtschreibkompetenzmodelle

Normorientierte Modelle zum Schriftspracherwerb folgen häufig der Perspektive „eines Schriftkundigen, der sprachliches Wissen schon erworben hat“ (Hinney, 2010, S.61). Kinder mit einem rudimentär ausgebildeten Sprachbewusstsein verlieren aber so bereits zu Beginn des Schriftspracherwerbs den Anschluss. Ein prominentes Beispiel für ein normbasiertes Kompetenzmodell stellen die Annahmen zum Schriftspracherwerb von May (2013) dar. Er geht von einem Stufenmodell des Rechtschreiblernens aus. Zentral ist die An-

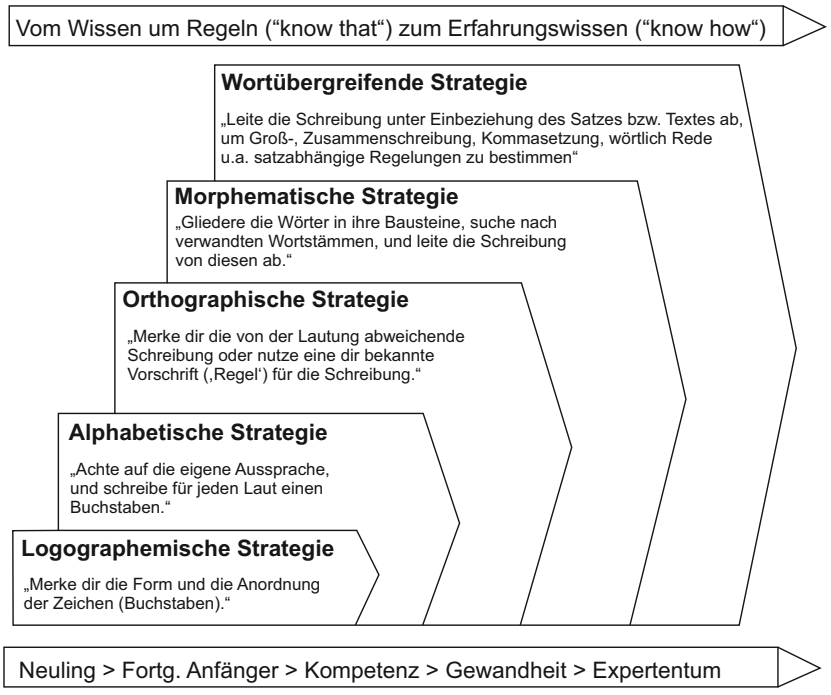

Abbildung 3.1: Entwicklungsstufen des Rechtschreibprozesses (nach May, 2013)

nahme, dass es grundlegende Strategien zur Erschreibung von Wörtern und Sätzen gibt. So können Regeln, die Schüler*innen beim Schreibenlernen entdecken und anwenden, bestimmten Prinzipien zugeordnet werden, die in der Schrift des Deutschen begründet sind. Neben dem Prinzip der Nachahmung Schriftkundiger und des Merkens von Buchstabenkombinationen (logographemisches Prinzip), kommt in diesem Ansatz besonders den alphabetischen und morphematischen Grundprinzipien eine wichtige Bedeutung zu. Ferner spielen orthografische Prinzipien eine wichtige Rolle, die auf der Grundlage des morphematischen Prinzips das alphabetische Prinzip beeinflussen. Rechtschreibstrategien und letztendlich eine umfassende Gesamtstrategie des Rechtschreibens entwickeln sich nach diesem Ansatz durch Hinweise im Unterricht und durch die eigenaktive Bildung von Regeln durch die Lernenden.Kinder wenden die handlungsleitenden Hinweise und (Selbst-)Instruktionen anfangs häufig unreflektiert an, ohne die zugrunde liegenden Regeln und Prinzipien zu verstehen, und machen dabei zwangsläufig Fehler. Mit zunehmender Erfahrung können sie die Regeln jedoch in unterschiedlichen Kontexten anwenden und ihr Verständnis vertiefen (May, 2009). Das „Lernen durch Instruktion“ wird demnach ergänzt und überformt durch ein „Lernen durch Tun“, und das Wissen der Lernenden schreitet vom Wissen über anzuwendende (Selbst-) Instruktionen, also Handlungsregeln (know that), zum intuitiven Wissen (know how) fort.

Durch die alphabetische, orthografische, morphematische und wortübergreifende Strategie, die im Folgenden näher erläutert werden, werden die grundlegenden Zugriffsweisen der Schüler*innen auf Schrift beschrieben, mit denen sich der jeweilige erreichte individuelle Lernstand erfassen lässt (May, 2013). Darüber hinaus spielen die Aspekte der „überflüssigen orthografischen Elemente“ und der „Oberzeichenfehler“ eine Rolle bei der Feststellung der individuellen Rechtschreibleistungen der Schüler*innen. Die Zugriffsweise der alphabetischen Strategie basiert auf der Analyse des eigenen Sprechens. Es geht also um die Fähigkeit, den Lautstrom der Wörter zu erschließen und diesen mit Hilfe von Buchstaben bzw. Buchstabenkombinationen schriftlich festhalten zu können (May, 2013). Die orthografische Strategie umfasst die Fähigkeit, Laut-Buchstaben-Zuordnungen unter Berücksichtigung spezifischer orthografischer Prinzipien und Regeln anpassen zu können. Zu diesen orthografischen Elementen gehören sowohl Merkelemente wie z.B. <Zahn>, <Vater> und <Hexe>, die sich die Lernenden einprägen müssen, als auch Regelelemente, wie z.B. <Koffer>,<stehen> und <Hand>, deren Verwendung hergeleitet werden kann (May, 2013). Die morphematische Strategie bezieht sich auf die Fähigkeit, bei der Schreibweise die morphematische Struktur der Wörter zu berücksichtigen. Hierbei werden sowohl die Wortart zur Ableitung der Groß-/Kleinschreibung als auch die Wortsemantik für die Zusammen- oder Getrenntschreibung und die Satzgrammatik berücksichtigt.

Die Anwendung der morphematischen Strategie setzt sowohl die Fähigkeit voraus, den jeweiligen Wortstamm zu erschließen, wie bei den Wörtern <Staubsauger> und <Räuber> (morphosemantische Bedeutungskompetenz), als auch die Fähigkeit, komplexe Wörter in ihre Bestandteile zu zerlegen, wie bei den Wörtern <Fahrrad> und <Geburtstag> (morphologische Strukturkompetenz) (May, 2013).

Die wortübergreifende Strategie beinhaltet die Fähigkeit, bei der Verfassung von Sätzen und Texten weitere sprachliche Aspekte zu berücksichtigen, wie beispielsweise die Zu-

sammenhänge zwischen Sätzen, die Verwendung von Satzzeichen wie Kommas oder die korrekte Anwendung von Konnektoren. Hierbei spielen auch die Wortart zur Herleitung von Groß- und Kleinschreibung sowie die Wortsemantik eine wichtige Rolle (May, 2013).

3.2.2 Sprachsystematisches Rechtschreibkompetenzmodell

Die sprachsystematische Sichtweise auf die Domäne Rechtschreibung lässt sich von der normorientierten Sichtweise unterscheiden, indem sie das Konstrukt der Rechtschreibung systematisiert und in einen Kern- und Peripheriebereich aufteilt und dadurch neue Perspektiven für den sprachlichen Anfangsunterricht eröffnet (vgl. Kap. 2.2) (Blatt & Pagel, 2009). Das Modell berücksichtigt die sprachlichen Voraussetzungen der Kinder und setzt anstelle der lautgetreuen Schreibung oder des Regellernens und Übens das Erkunden und Verstehen der Schriftstrukturen als Grundlage des Rechtschreiberwerbs (Blatt et al., 2015; Blatt, Prosch & Frahm, 2016). Sie umfasst nicht nur eine Zusammenstellung von orthografischen Regeln auf den Grundlagen der amtlichen Rechtschreibung, sondern auch eine Systematik der Orthografie, die eine Einsicht in die Schriftstruktur der Wörter ermöglicht und die Regularitäten der Wortschreibung erklärbar macht (Eisenberg, 2016). Die sprachsystematische Sichtweise wendet sich vom vorherrschenden Konzept der lautgetreuen Schreibung zum Schriftspracherwerb ab, in dem das Geschriebene als Abbild der gesprochenen Sprache gilt. Der Rechtschreiberwerbsprozess dient nicht nur zur Erfüllung gesellschaftlicher Normen, vielmehr steht das hohe Lernpotenzial dieses Prozesses für die mündliche und schriftliche Sprachkompetenz im Fokus (Eisenberg & Fuhrhop, 2007). Der sprachsystematische Rechtschreibtest (SRT) wurde von Inge Blatt und Andreas Voss auf Basis der Ergebnisse der graphematischen Forschung Eisenbergs (1995) und Didaktik Hinney's (1997) entwickelt und im Rahmen der Ergänzungsstudie „Rechtschreibung zur Internationalen Grundschulleseuntersuchung“ (IGLU-E-Studie 2006) erstmals eingesetzt (Blatt et al., 2015). Rechtschreibkompetenz wird im SRT als ein komplexes kognitives Konstrukt erfasst, differenziell überprüft und als Ausdruck einer jeweils spezifischen Kompetenz im Sinne von Hartig und Klieme (2006) verstanden (Blatt et al., 2015, S.229). Blatt & Frahm (2013, S.18) gehen davon aus, dass sich die Rechtschreibfähigkeit „abhängig von den individuellen Lernvoraussetzungen nach und nach entwickeln und in ein komplexes mentales Modell der Rechtschreibkompetenz integrieren“. Die Automatisierung der Teilkompetenzen ist die Voraussetzung für die Weiterentwicklung von Rechtschreibkompetenz. Im Gegensatz zu anderen Entwicklungsmodellen geht es nicht von einer strikten Abfolge bestimmter Phasen aus, sondern benennt einzelne Teilkompetenzen, die der Rechtschreibung zugeordnet werden können. Die Rechtschreibkompetenz lässt sich demnach in fünf Prinzipien ausdifferenzieren, denen jeweils Teilkompetenzen zugeordnet werden, die unterschiedliche Anforderungen an den Schreibenden darstellen (vgl. Abb. 6.1).

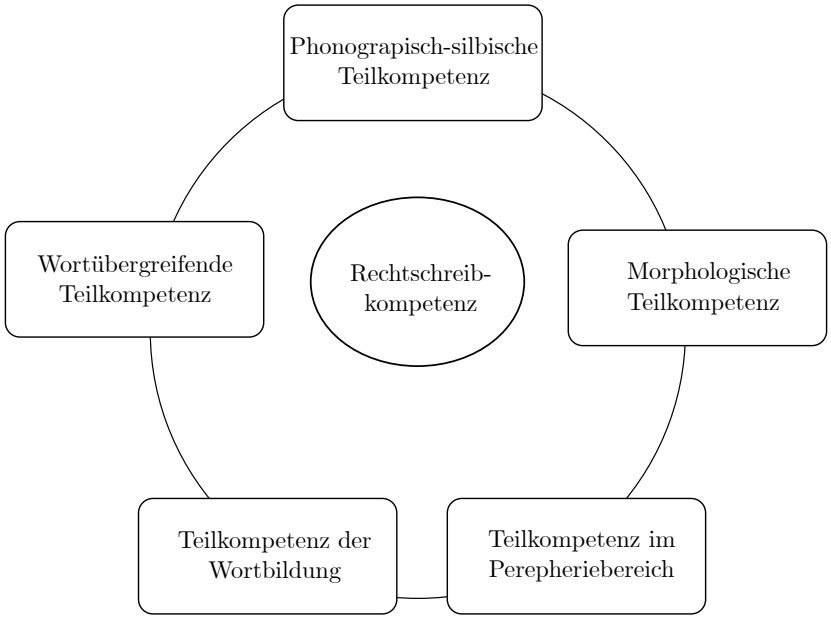

Abbildung 3.2: Rahmenkonzeption zum sprachsystematischen Rechtschreibtest (Blatt et al., 2015, S.237)

Das phonographische, silbische und morphologische Prinzip Eisenbergs (1995) stellt die theoretische Grundlage für die Teilkompetenzen im Kernbereich des Rechtschreibkompetenzmodells dar, wobei das phonographische und silbische Prinzip in einer Teilkompetenz zusammengefasst wird. Im phonographisch-silbischen Prinzip geht es um die Anforderung, den regelhaften Zusammenhang zwischen gesprochener und geschriebener Sprache unter Beachtung der silbenstrukturellen Informationen (z.B. Vokalquantität, Betonung, Silbenschnitt) zu kennen. Der morphologische Bereich des deutschen Schriftsystems wird zum einem durch das morphologische Prinzip im Kernbereich und zum anderen durch das Prinzip der Wortbildung repräsentiert. Dies ist der Tatsache geschuldet, dass mit der Beherrschung dieser beiden Bereiche unterschiedliche Fähigkeiten gefordert sind (Blatt & Frahm, 2013). Im Rahmen des morphologischen Prinzips werden die Struktureinheiten von einsilbigen und flektierten Wörtern und die Umlautschreibung aufgrund der morphologischen Konstanz untersucht und Flexionsmorpheme richtig angewendet. Das Prinzip der Wortbildung umfasst Präfixe und Suffixe und Komposita. Die Groß- und Kleinschreibung wird mittels des wortübergreifenden Prinzips repräsentiert (Blatt & Frahm, 2013).

Zum sprachsystematischen Rechtschreibkompetenzmodell liegt eine umfassende Begleitforschung vor. In den Studien IGLU-E Vorstudie, NEPS und HeLp wurde das sprachsystematische Rechtschreibkompetenzmodell auf seine Strukturen hin überprüft und mehrfach ein fünfstufiges Modell für die vierte und fünfte Klasse nachgewiesen, das die differenziellen Teilkompetenzen abbildet (Blatt, Prosch & Lorenz, 2016; Jarsinski, 2014; Naujokat, 2015;Voss et al., 2007). Die Ergebnisse der IGLU-Vorstudie 2005 (N=486) und der NEPS-Großpilotstudie 2016 (N=492) mit Schüler*innen der vierten Klasse weisen darauf hin, dass sich mittels dieses Kompetenzmodells Unterschiede in Kompetenzstrukturen von

Leistungsgruppen identifizieren lassen. Im Rahmen einer empirischen Studie zu Mode-Effekten im Kontext des Nationalen Bildungspanels wurde zudem ein computerbasierter Test auf Basis des sprachsystematischen Rechtschreibtests zur Testung der Rechtschreibleistung in Klasse fünf entwickelt (Frahm, 2013). Der SRT-Editor ist ein Programm zur automatisierten Codierung des Sprachsystematischen Rechtschreibtests, identifiziert automatisch die Struktureinheiten der Schreibungen der Schüler*innen und codiert diese. Basis bildet das Regelwerk von Frahm (2013), anhand dessen die Wortstruktur der Wörter auf Grundlage des sprachsystematischen Kompetenzmodells festgelegt werden können. Die Ergebnisse der Studie zeigen, dass sich auf der Grundlage des Regelwerks eine automatisierte Codierung entwickeln lässt, die im Vergleich zur manuellen Codierung zuverlässiger ist (Frahm, 2013). Dies stellt die Grundlage für ein formatives Assessment auf Basis des sprachsystematischen Rechtschreibkompetenzmodells dar, da nur durch eine ökonomische und automatische Codierung eine sofortige Ergebnisauswertung möglich ist (Frahm, 2013). Der SRT-Editor ist jedoch nur für Wissenschaftler*innen zugänglich, die mit Datensätzen des Nationalen Bildungspanels zur Rechtschreibkompetenzentwicklung forschen.

3.3 Rechtschreibkompetenz in den Bildungsstandards

In diesem Abschnitt werden die Bedeutung der Bildungsstandards für das Fach Deutsch in der Primarstufe und die Erkenntnisse der empirischen Bildungsforschung zur Kompetenzmodellierung für die Testkonstruktion dargestellt.

Für den Deutschunterricht der Primarstufe hat die ständige Konferenz der Kultusminister der Länder (Kultusminister Konferenz, 2005) Bildungsstandards verabschiedet. Seit dem Schuljahr 2005/2006 bilden sie eine verbindliche Grundlage für die Lehrplan- und Schulentwicklung in allen Bundesländern. Die Bildungsstandards beziehen sich auf die Kernbereiche des Fachs, benennen Grundprinzipien und beschreiben kumulative Lernprozesse und normativ gesetzte Kompetenzerwartungen, die zu einem bestimmten Zeitpunkt von allen Schüler*innen als Regelstandard erreicht werden sollten (Stanat et al., 2017). Nachstehend werden die Bildungsstandards mit Blick auf die für diese Arbeit relevanten Bereiche der Schriftlichkeitsforschung und des Rechtschreibens vorgestellt. Die Bildungsstandards im Fach Deutsch für die Jahrgangsstufe vier in der Grundschule (Kultusminister Konferenz, 2005) beziehen sich auf eine vorschulisch erworbene Sprachhandlungskompetenz, die „in den Bereichen des Sprechens und Zuhörens, des Schreibens, des Lesens und Umgehens mit Texten und Medien sowie des Untersuchens von Sprache und Sprachgebrauch“ (Kultusminister Konferenz, 2005, S.7) erweitert werden sollen. Alle angestrebten Kompetenzen sind als „Können-Beschreibungen“ (Can-do Statements) angegeben und nicht explizit definiert. Damit spiegeln sie die in der fachdidaktischen und sprachwissenschaftlichen Literatur übliche Einteilung sprachlicher Kompetenzen in Lesen, Hören, Schreiben und Sprechen wider (Felder, 2003). Anhand der Sprachmodalität lassen sich mündliche Kompetenzen (Sprechen, Hören, zuhören) von schriftsprachlichen Kompetenzen (Schreiben, Lesen) unterscheiden.

Sie umfassen jeweils „Methoden und Arbeitstechniken“:

- Sprechen und Zuhören: zu anderen sprechen, verstehend zuhören, Gespräche führen, szenisch spielen, über lernen sprechen
- Schreiben: über Schreibfertigkeiten verfügen, richtig schreiben, Texte planen, Texte schreiben, Texte überarbeiten
- Lesen – mit Texten und Medien umgehen: über Lesefähigkeiten verfügen, über Leseerfahrungen verfügen, Texte erschließen, Texte präsentieren
- Sprache und Sprachgebrauch untersuchen: grundlegende sprachliche Strukturen und Begriffe kennen, sprachliche Verständigung untersuchen, an Wörtern, Sätzen, Texten arbeiten, Gemeinsamkeiten und Unterschiede von Sprache entdecken (Böhme et al., 2017, S.8)

Zwei Kompetenzbereiche, nämlich „Schreiben“ und „Sprache und Sprachgebrauch untersuchen“ sind dem „richtig schreiben“ und der Arbeit „an Wörtern, Sätzen, Texten“ gewidmet. Für den Kompetenzbereich „Schreiben“ führen die Bildungsstandards aus:

> „Die Kinder verfügen über grundlegende Rechtschreibstrategien. Sie können lautentsprechend verschriften und berücksichtigen orthografische und morphematische Regelungen und grammatisches Wissen. Sie haben erste Einsichten in die Prinzipien der Rechtschreibung gewonnen. Sie erproben und vergleichen Schreibweisen und denken über sie nach. Sie gelangen durch Vergleichen, Nachschlagen im Wörterbuch und Anwenden von Regeln zur richtigen Schreibweise. Sie entwickeln Rechtschreibgespür und Selbstverantwortung ihren Texten gegenüber“ (Kultusminister Konferenz, 2005, S.8).

Im Kompetenzbereich „Sprache und Sprachgebrauch untersuchen“ heißt es:

> „Anknüpfend an ihre Spracherfahrungen entwickeln die Kinder ihr Sprachgefühl weiter und gehen bewusster mit Sprache um. In altersgemäßen, lebensnahen Sprach- und Kommunikationssituationen erfahren und untersuchen die Kinder die Sprache in ihren Verwendungszusammenhängen und gehen dabei auf die inhaltliche Dimension und die Leistung von Wörtern, Sätzen und Texten ein“ (Kultusminister Konferenz, 2005, S.9).

Für den Kompetenzbereich „Schreiben“ mit dem Schwerpunkt Rechtschreiben werden die folgenden Standards benannt (Kultusminister Konferenz, 2005, S.10-11):

- „geübte, rechtschreibwichtige Wörter normgerecht schreiben,
- Rechtschreibstrategien verwenden: Mitsprechen, Ableiten, Einprägen,
- Zeichensetzung beachten: Punkt, Fragezeichen, Ausrufezeichen, Zeichen bei wörtlicher Rede,
- über Fehlersensibilität und Rechtschreibgespür verfügen
- Rechtschreibhilfen verwenden

  - Wörterbuch nutzen,
  - Rechtschreibhilfen des Computers kritisch nutzen

- Arbeitstechniken nutzen
  - methodisch sinnvoll abschreiben
  - Übungsformen selbstständig nutzen,
  - Texte auf orthographische Richtigkeit überprüfen und korrigieren".

Zu den Bildungsstandards im Fach Deutsch für die Grundschule liegt eine intensive Auseinandersetzung seitens der Bildungsforschung und der Fachdidaktik vor. So kritisiert Köller (2009) , dass einige Merkmale wie Kumulativität, Verständlichkeit, Realisierbarkeit sowie Messbarkeit nur zum Teil erfüllt seien. Hinney (2010, S.63) weist darauf hin, dass die Bildungsstandards zu schnell im Anschluss an die Veröffentlichung der Bildungsexpertise (Klieme et al., 2003) und ohne Vorliegen wissenschaftlicher Kompetenzmodelle veröffentlicht wurden. Offengeblieben ist die Frage, warum Rechtschreiben und Textschreiben in einem eigenen Kompetenzbereich „Schreiben" zusammengefasst sind und die Orthografie nicht dem Bereich „Sprache und Sprachgebrauch untersuchen" zugeordnet wird (Bremerich-Vos, 2009, S.199). Ferner ist empirisch ungeklärt „inwieweit es sich bei den lediglich additiv benannten Teilkompetenzen um analytisch trennbare Dimensionen handelt" (Granzer et al., 2008, S.13) oder ob und worin ein Zusammenhang beim Bereich „Schreiben" benannten Kompetenzen besteht.

Hinney (2010, S.64) attestiert zudem der „statisch-linearen" Beschreibung der Kompetenzen in Bezug auf lautorientiertes, orthografisches und morphematisches Schreiben und deren jeweiligen Strategien eine orthografietheoretisch einseitige Orientierung. Eine Unterteilung der Rechtschreibkompetenz in „mitsprechen", „ableiten" und „einprägen" sieht sie kritisch. So sei diese Strategie des Mitsprechens nicht für Kinder geeignet, die Schwierigkeiten beim deutlichen Sprechen haben. Für diese Kinder sei es schwer „aus der verkürzten und verschliffenen Umgangssprache (z.B. „fümf") [...] zur angemessenen Lautung und richtigen Schreibung" zu kommen (Hinney, 2014, S.459). Als positive Aspekte der Kompetenzdefinition beim Rechtschreiben hebt sie jedoch den kognitiven Zugang, das problemlösende Vorgehen sowie die Förderung des orthografischen Zweifelns und des eigenverantwortlichen orthografischen Überarbeitens eines Textes hervor (Hinney, 2014, S.459).

Aufgrund solcher orthografietheoretischer und – didaktischer Differenzen (2.1.4) und aufgrund von Ergebnissen empirischer Studien des Instituts zur Qualitätsentwicklung im Bildungswesen (IQB) wird Orthografie in der Diagnostik als eigener Teilbereich behandelt und separat getestet (Böhme et al., 2017). Auch aus messtheoretischen Erwägungen liegt es nahe, Orthografie als eigenen Teilbereich zu betrachten.

3.4 Kompetenzbegriff der empirischen Bildungsforschung

Während herkömmliche Lehrpläne und Unterrichtsmaterialien die Zielsetzungen für das jeweilige Unterrichtsfach erörtern, mit welchen Inhalten und Methoden diese erreicht werden sollen, richtete sich der Blick mit der empirischen Bildungsforschung auch darauf, was die Schüler*innen tatsächlich können und welche Fähigkeiten sie als Resultat von Bildungsprozessen erworben haben (Klieme et al., 2003; Klieme et al., 2007). Zentraler Begriff zur Erfassung von Können und Fähigkeiten ist der Kompetenzbegriff, der schon vor einigen Jahrzehnten mit der Theorie der Sprachkompetenz von Noam Chomsky (1968) Verwendung fand. Seit den internationalen Schulleistungsuntersuchungen wie TIMSS und PISA hat der Kompetenzbegriff auch Einzug in die pädagogische Diskussion sowie in Lehrpläne und KMK-Bildungsstandards gehalten. Insbesondere die Ergebnisse dieser Studien zeigen die Diskrepanz zwischen den Zielen und den tatsächlich erreichten Kompetenzen der Schüler*innen.

Noch immer ist der Kompetenzbegriff Gegenstand kontroverser wissenschaftlicher Diskussionen, wird in der Praxis vielfältig verwendet und gilt manchen als ein pädagogischer Modebegriff (Prenzel et al., 2007). Die folgende Definition des Kompetenzbegriffs von Weinert (2001) bildet eine wichtige Grundlage bei der Bestimmung des Begriffs. Kompetenzen sind:

> „die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können“ (Weinert, 2001, S.27).

Hartig und Klieme (2006) definieren Kompetenzen, die zur Bewältigung verschiedener Situationen beitragen „als kontextspezifische kognitive Leistungsdisposition, die sich funktional auf bestimmte Klassen von Situationen und Anforderungen beziehen. Diese spezifischen Leistungsdispositionen lassen sich auch als Kenntnisse, Fertigkeiten oder Routinen charakterisieren“. Zur individuellen Kompetenz einer Person tragen miteinander vernetzte Aspekte wie Wissen, Fähigkeit, Können, Handeln, Erfahrungen und Motivation bei, die sich bei einer tatsächlich erbrachten Leistung zeigen (Klieme et al., 2003, S.72). Diesen allgemeinen Kompetenzbegriff gilt es jeweils domänenspezifisch im Kontext verschiedener Unterrichtsfächer, Lernbereiche oder Aufgabenfelder zu spezifizieren. So hat Ossner (2006) beispielsweise für den Lernbereich Rechtschreibung ein Strukturmodell entwickelt, das Kompetenz entsprechend lernpsychologischer Ansätze nach vier verschiedenen Wissensbereichen wie deklaratives Wissen, Problemlösewissen, prozedurales und metakognitives Wissen differenziert.

Wissensart...	**... bezogen auf die Rechtschreibkompetenz**
deklaratives Wissen	stoffliches Wissen wie z.B. Definitionen und Rechtschreibregeln kennen Merksatz: Deklaratives Wissen entsteht durch Faktenlernen.
Problemlösungswissen	Strategien zur Herleitung der richtigen Schreibung kennen, z.B. morphematische Ableitung Merksatz: Problemlösungswissen zeigt sich in der intelligenten Anwendung von Methoden zur Erkenntnisgewinnung.
prozedurales Wissen	die Beherschung der Orthografie, wenn sie keine besondere Aufmerksamkeit mehr braucht Merksatz: Prozedurales Wissen entsteht durch Üben und zeigt sich vor allem im automatisierten Können.[...] Es ist ein implizites Wissen.
metakognitives Wissen	Kenntnis seiner eigenen Fähigkeiten und Grenzen in der Orthografie, beispielsweise bezogen auf Fehlerschwerpunkte oder nützliche Lernstrategien Merksatz: Metakognitives Wissen ist eng an Reflexion geknüpft.

Tabelle 3.1: Wissensarten in Anlehnung an Ossners Strukturmodell (2006)

Je enger eine Kompetenzdomäne bestimmt wird, desto bereichsspezifischer sind die erforderlichen Kompetenzen zu bestimmen. Während die umfassende Definition des Kompetenzbegriffs Weinerts (2001) sowohl kognitive als auch motivationale (z.B. Fachinteresse) und volitionale Faktoren (z.B. Schulangst) umfasst, findet in internationalen Schulleistungsstudien wie TIMSS und PISA eine Begrenzung auf domänenspezifische kognitive Leistungsdispositionen statt. Dies hat theoretische, aber nicht zuletzt auch pragmatische Gründe, da der umfassende Kompetenzbegriff hohe Anforderungen an die Operationalisierung und Messung von Kompetenzen stellt. Auch die vorliegende Arbeit fokussiert auf die kognitiven Aspekte des Kompetenzbegriffs. Dementsprechend sind Kompetenzen „kontextspezifische kognitive Leistungsdispositionen, die sich funktional auf Situationen und Anforderungen in bestimmten Domänen beziehen“ (Klieme & Leutner, 2006, S.879). Was in einer Domäne als kompetentes Handeln gilt, unterliegt Veränderungen und hängt von unterschiedlichen fachdidaktischen Konzepten und Kompetenzmodellen ab.

In der empirischen Bildungsforschung stellen Kompetenzen

> „theoretische Konstrukte dar. Aus der inhaltlichen Definition eines Konstrukts leitet sich im wissenschaftlichen Kontext ab, wie es in einer empirischen Untersuchung operationalisiert werden sollte, d. h. mit welchen Methoden und Instrumenten eine *Messung* erfolgen sollte. Wissenschaftliche Hypothesen werden auf Basis empirischer Daten beurteilt, die auf solchen Operationalisierungen theoretischer Konstrukte basieren“ (Hartig, 2008, S.16).

Um den Anforderungsbereich zu strukturieren und für die Messung zu erschließen, legt man ein Kompetenzmodell zugrunde, in das konkrete Aufgaben eingeordnet werden können. Mittels solcher Aufgaben wird eine objektive, reliable und valide Aussage über die Kompetenzen der Schüler*innen in der untersuchten Domäne erwartet.

In Kompetenzmodellen geht es darum, den

> „Inhaltsbereich sinnvoll zu strukturieren, so dass einerseits der Inhaltsbereich in Kompetenzbereiche untergliedert wird und andererseits erkennbar wird, welche Anforderungen ein Individuum in den Kompetenzbereichen erfüllen kann. Dazu werden die Anforderungen gestuft, und es wird versucht, eine Verbindung zwischen den beschriebenen Anforderungen auf der einen Seite und den kognitiven Prozessen des Individuums auf der anderen Seite herzustellen“ (Münzer, 2012, S.3).

Die Modellierung und Messung von Kompetenzen gehört zu einem interdisziplinären Forschungsgebiet, an dem Wissenschaftler*innen aus den Bereichen Erziehungswissenschaft, Psychologie und verschiedener Fachdidaktiken beteiligt sind (Klieme et al., 2007).

3.5 Zusammenfassung

Schüler*innen haben zu Schulbeginn unterschiedliche Lernausgangslagen und Vorerfahrungen mit der Schriftsprache, die von der familiären Literalität und sozialem Status beeinflusst werden. Es ist wichtig, dass Schüler*innen im Unterricht Rechtschreibung als systematisch und erlernbar vermittelt bekommen und die individuellen Lernvoraussetzungen berücksichtigt werden (Weinhold et al., 2020). Die Bildungsstandards bieten nur eine grobe Orientierung für die Definition des zu erfassenden Lernbereichs und die theoretische Fundierung für die Konstruktion des Rechtschreibkompetenz-Messverfahrens. Die empirische Bildungsforschung hingegen misst mit Hilfe von Kompetenzmodellen die tatsächlichen Lernergebnisse von Bildungsprozessen der Schüler*innen. Kompetenzmodelle sind für die Strukturierung und Erschließung des zu messenden Anforderungsbereichs für diagnostische Testverfahren zentral, da konkrete Aufgaben in sie eingeordnet werden können. Diese Vorgehensweise ist für die Testkonstruktion wichtig, da sie es ermöglicht, Aufgaben zu definieren, die eine objektive, reliable und valide Aussage über die Kompetenzen der Schüler*innen in der untersuchten Domäne erlauben. Was in einer Domäne, wie im Kontext der vorliegenden Arbeit dem Schriftspracherwerb, als kompetentes Handeln gilt, unterliegt Veränderungen und hängt, wie gezeigt, von unterschiedlichen fachdidaktisch orientierten Konzepten und Kompetenzmodellen ab. Für eine theoriegeleitete Auswahl eines Rechtschreibkompetenzmodells zur Konstruktion eines zuverlässigen Rechtschreibkompetenz-Messverfahrens ist es notwendig, vorhandene Modelle hinsichtlich ihrer unterschiedlichen (Lern-)psychologischen und fachlichen Zugänge zur Domäne Rechtschreibung zu analysieren.

In der Orthografieforschung wird zunehmend Kritik an einem normorientierten Verständnis vom Schriftspracherwerb geäußert (z.B. Blatt et al., 2021; Hinney, 2010; Jagemann

& Weinhold, 2018; Weinhold et al., 2020) und ein Paradigmenwechsel hin zu einer systemorientierten Modellierung des Konstrukts Rechtschreibung gefordert (Blatt, Prosch & Lorenz, 2016). Ein besonders geeigneter Ansatz ist hierbei die sprachsystematische Sichtweise, die auf Erkenntnissen der graphematischen Forschung basiert und Schüler*innen das Verständnis für die Struktur des Schriftsystems und dessen Regularitäten vermittelt. Der Diskussionsstand zeigt auch, dass die normbasierten Stufenmodelle als Grundlage für ein didaktisches Handeln den heterogenen Entwicklungsverläufen der Kinder beim Schriftspracherwerb nicht gerecht werden können und die sprachsystematische Sichtweise auf den Lerngegenstand Rechtschreibung große Potenziale bietet. Nur wenn die individuellen Lernverläufe erfasst werden, kann darauf aufbauend ein differenzierter Unterricht realisiert werden. Das strukturorientierte und prozessuale, sprachsystematische Rechtschreibkompetenzmodell eröffnet im Spannungsfeld zwischen verallgemeinernden Modellen und individualisierten Betrachtung (Bulut, 2018) große Potenziale für eine individualisierte Lernverlaufsdiagnostik. Das Modell berücksichtigt die sprachlichen Voraussetzungen der Schüler*innen und setzt anstelle des lautgetreuen Schreibens oder des Regellernens das Erkunden und Verstehen von Schriftstrukturen für den Erwerb von Rechtschreibkompetenz (Blatt et al., 2015). Schüler*innen mit Problemen beim Schriftspracherwerb profitieren insbesondere von einem Unterricht, der die Struktur der Schriftsprache vermittelt, die unter der Annahme der Interdependenzhypothese in der Graphematik vertreten wird (Bangel & Müller, 2018; Bulut, 2018; Hein & Blatt, 2016). Die Verwendung eines sprachsystematischen Ansatzes basierend auf Erkenntnissen der graphematischen Forschung gilt als zentrale theoretische Grundlage zur Konstruktion eines zuverlässigen Rechtschreibkompetenz-Messverfahrens und zur Vermittlung von Rechtschreibkompetenz in der Schule.

Inwiefern auf der theoretischen Grundlage des sprachsystematischen Ansatzes eine engmaschige Leistungsmessung konzipiert werden kann, um Rück- und Fortschritte sowie Stagnationen in den unterschiedlichen Bereichen der Orthografie sichtbar machen zu können, wird in den folgenden Kapiteln erläutert.

4 Lernverlaufsdiagnostik von Rechtschreibkompetenz

Die Heterogenität der Lernausgangslagen und Lernprozesse beim Erwerb der Rechtschreibkompetenz sowie die diagnostischen Fähigkeiten der Lehrkräfte erfordern eine regelmäßige Lernstandserhebung mit standardisierten und reliablen Testverfahren. Das Konzept der Lernverlaufsdiagnostik gilt vor diesem Hintergrund als das wichtigste Instrument für erfolgreiche Lernprozesse und zur Verbesserung von Bildungsqualität. In diesem Kapitel werden Fragen nach den Zielen, Methoden und Gelingensbedingungen der Lernverlaufsdiagnostik geklärt. Daran schließt sich eine Analyse der bisher vorliegenden Instrumente zur Lernverlaufsdiagnostik im Bereich Rechtschreibung an (vgl. Kap. 4.2). Anschließend werden die besonderen methodischen Herausforderungen diskutiert, die mit der Entwicklung von Instrumenten zur Lernverlaufsdiagnostik verbunden sind (vgl. Kap. 4.3, Kap. 4.4).

4.1 Lernverlaufsdiagnostik

Bevor die Ziele und Methoden und Gelingensbedingungen der Lernverlaufsdiagnostik erläutert werden, folgt zunächst die Einordnung des Ansatzes der Lernverlaufsdiagnostik in den übergreifenden Kontext von „Diagnostik“. Angesichts vielfältiger Ansätze und nicht eindeutig verwendeter Begrifflichkeiten wie z.B. summative und formative Evaluation, summative und formative Diagnostik und summatives und formatives Assessement, ist eine Beschreibung notwendig.

4.1.1 Begriffliche Einordnung

Die Diagnostik lässt sich nach den jeweiligen Disziplinen und spezifischen Anwendungsfeldern wie der psychologischen Diagnostik, pädagogischen oder pädagogisch-psychologischen Diagnostik, Schulleistungsdiagnostik und formativer Diagnostik bzw. Lernverlaufsdiagnostik einordnen und beschreiben. Sie ist ein Teilgebiet verschiedener Disziplinen wie Psychologie, Pädagogik und Inklusionspädagogik. Es gibt sowohl Gemeinsamkeiten und deutliche Überschneidungen oder äquivalente Verwendungen wie z.B. bei Begriffen wie Evaluation und assessment als auch nicht klar definierte bzw. abgegrenzte Begriffe.

Psychologische Diagnostik

Eid und Petermann (2006, S.16) legen einen umfassenden Definitionsvorschlag für die psychologische Diagnostik vor:

> „Die Inhalte und Methoden der Psychologischen Diagnostik beziehen sich auf die regelgeleitete Sammlung und Verarbeitung von gezielt erhobenen Informationen, die für das Verständnis menschlichen Verhaltens und Erlebens bedeutsam sind. Aus den gewonnenen Informationen sollen Fragestellungen (eines Auftragsgebers) bearbeitet und Entscheidungen getroffen werden. Die Prinzipien der Entscheidungsfindung müssen wissenschaftlichen Kriterien entsprechen. [...] Die Fragestellungen der Psychologischen Diagnostik können sich dabei auf die
>
> - Beschreibung und
> - Klassifikation,
> - Erklärung,
> - Vorhersage (Prognose) und
> - Evaluation von Zuständen und/oder Verläufen beziehen."

Fisseni (2004, S.4) verweist auf einen Zusammenhang von Diagnostik und Intervention: „Diagnostik soll zur Intervention führen, Intervention setzt Diagnostik voraus." Da sich die pädagogische Diagnostik, auch pädagogisch-psychologische Diagnostik genannt, der Forschungsmethoden der psychologischen Diagnostik bedient (Ingenkamp & Lissmann, 2008), ist die folgenden von Petermann und Wirtz (2023, S.1) vorgenommenen Gegenüberstellung der Status- und Veränderungsdiagnostik für den Kontext der vorliegenden Arbeit relevant:

> „Status- vs. Veränderungsdiagnostik: Die Statusdiagnostik zielt darauf ab, den Ist-Zustand in Bezug auf die für die Problemstellung zentralen Merkmale zu beschreiben. Neben der deskriptiven Darstellung kann es u.a. das Ziel sein, (a) zukünftige Entwicklungen zu prognostizieren (z. B. Schuleignung, Berufseignung) oder (b) Ursachen oder zugrunde liegende Merkmalsausprägungen (z. B. Wahrnehmungseinschränkungen bei manifest diagnostizierten Einschränkungen der Leseleistung) erkennen zu können. Bei der Verlaufsdiagnostik werden Merkmale zu mehreren Messzeitpunkten erhoben, um z. B. natürliche Prozesse oder Effekte von Interventionen zu dokumentieren."

Pädagogische Diagnostik

Während sich die psychologische Diagnostik primär auf die Testung zur Klassifizierung und Typologisierung menschlichen Verhaltens bezieht, umfasst pädagogische Diagnostik:

> „diagnostischen Tätigkeiten, durch die bei einzelnen Lernenden und den in einer Gruppe Lernenden Voraussetzungen und Bedingungen planmäßiger Lehr- und Lernprozesse ermittelt, Lernprozesse analysiert und Lernergebnisse festgestellt werden, um individuelles Lernen zu optimieren. Zur Pädagogischen Diagnostik gehören ferner die diagnostischen Tätigkeiten, die die Zuweisung zu Lerngruppen oder zu individuellen Förderungsprogrammen ermöglichen sowie die mehr gesellschaftlich verankerten Aufgaben der Steuerung des Bildungsnachwuchses oder der Einteilung von Qualifikationen zum Ziel haben“ (Ingenkamp & Lissmann, 2008, S.13).

Die Bedeutung der pädagogischen Diagnostik hat sich aufgrund der verstärkten Diskussion um die Qualitätssicherung von Bildungsprozessen sowie um die bestmögliche Förderung von Schüler*innen mit unterschiedlichen Leistungsniveaus erhöht (Eid & Petermann, 2006).

Schulleistungsdiagnostik

Die Schulleistungsdiagnostik widmet sich der „systematische[n] Beschreibung und anschließenden Bewertung eines aktuellen Wissens- oder Fähigkeitsstatus von Lernenden bezüglich eines umschriebenen Inhaltsbereichs“ (Langfeldt & Imhof, 1999, S.281). Es handelt sich dabei in der Regel um standardisierte Testverfahren, die die Kompetenzen in ausgewählten Inhaltsbereichen d.h. das Leistungsvermögen zu einem bestimmten Zeitpunkt erfassen. Mittels statusdiagnostischer Verfahren erfolgt eine summative Leistungsbeurteilung oder „summative Diagnostik“, die meist die Ergebnisse der Evaluation eines längerfristigen Lernvorgangs umfassen (Hasselhorn et al., 2014).

Evaluation

Der Begriff Evaluation ist ebenfalls komplex und steht im Kontext eigener theoretischer und forschungsmethodologischer Erwägungen (Döring & Bortz, 2016). Nach einer umfassenden Analyse vorliegender wissenschaftlicher Definitionen beschreibt List-Ivankovic (2013, S.14) den Begriff wie folgt:

> „Evaluation richtet sich auf einen bestimmten Gegenstand, der durch die Anwendung sozial-wissenschaftlicher Methoden untersucht wird. Primäres Ziel einer Evaluation ist das Bewerten nach Kriterien, die nachvollziehbar und begründet entwickelt, dargestellt und angewendet werden. Die Evaluationsergebnisse sollen eine Grundlage für Planungen und Entscheidungen bieten und zur Verbesserung der Praxis durch die Bereitstellung von Handlungswissen beitragen“.

Summative Evaluation

Je nach Zeitpunkt und Ziel der Evaluation ist es üblich, zwischen summativer und formativer Evaluation zu unterscheiden. Diese Unterscheidung geht auf Scriven (1967, 1991) zurück:

> „Summative evaluation of a program (or other evaluand) is conducted after completion of the program (for ongoing programs, that means after stabilization) and for the benefit of some external audience or decision-maker (for example, funding agency, oversight office, historian, or future possible users), though it may be done by either internal or external evaluators or a mixture“ (Scriven, 1991, S.340).

Bei einer summativen Evaluation erfolgt eine Kontrolle oder Bewertung eines Evaluationsgegenstandes in Hinblick auf die Qualität oder Wirksamkeit (z.B. eines Projektes, Programms) nach dessen Fertigstellung oder Abschluss. Die summative Evaluation ist ergebnisorientiert und nicht prozessorientiert, ihre Ergebnisse richten sich an die Öffentlichkeit oder an Entscheidungsträger.

Formative Evaluation

Unter einer formativen Evaluation versteht Scriven:

> „Formative evaluation is contrasted with summative evaluation. It is typically conducted during the development or improvement of a program or product (or person, and so on) and it is conducted, often more than once, for the in-house staff of the program with the intend to improve “ (Scriven, 1991, S.168-169).

Die formative Evaluation (auch Prozessevaluation genannt) zielt darauf ab, einen Evaluationsgegenstand (z.B. Curriculum, Projekt, Bildungsprodukt) zu optimieren, während er sich noch in der Entwicklung befindet. Sie ist prozessorientiert, d.h. die Ergebnisse können für die Gestaltung und Optimierung des Evaluationsgegenstandes genutzt werden. Formative Evaluationen können auch mehrmals während eines Entwicklungsprozesses durchgeführt werden. Scriven (1991) weist auch auf die Funktion formativer Evaluation hin, mögliche Schwachstellen frühzeitig erkennen und Maßnahmen für deren Abhilfe ergreifen zu können. „... it is helpful to keep in mind that one of the most useful kinds of formative evaluation is early-warning summative“ (Scriven, 1991, S.169).

Die von Scriven (1967) eingeführte Unterscheidung zwischen summativer und formativer Evaluation gehört inzwischen zum „Allgemeingut“ (Klauer, 2014), wird äquivalent mit summativer und formativer Diagnostik oder zur Unterscheidung von summativen und formativen assessment verwandt. Gemeinsam ist Diagnostik und assessment, dass sie an einer Bestandsaufnahme ansetzen, „die auf der Ebene des Individuums als *Diagnostik*, auf der Aggregatebene von Schulklassen, Schulen, Institutionen und auf der Ebene des Bildungssystems als *Assessment* bezeichnet wird ...“ (Leutner, 2013, S.18).

Black und Wiliam (1998, S.141) verstehen unter assessment und insbesondere unter den Begriffen formative assessment:

> „all those activities undertaken by teachers, and by their students in assessing themselves, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged. Such assessment becomes 'formative assessment' when the evidence is actually used to adapt the teaching work to meet the needs."

Klauer (2014, S.2-3) weist in einem Überblick über die Entwicklung dieser Begrifflichkeiten darauf hin, dass „eine große Konfusion um die Unterscheidung von „formativ" und „summativ" festgestellt [wurde], weil sie zwar allgemein akzeptiert und gebraucht, aber vielfach beliebig umgedeutet wurde."

Auch in der bundesdeutschen Diskussion tauchen unterschiedliche Begrifflichkeiten auf. So klärt beispielsweise Walter (2014) in seinem Beitrag zu einer formativ orientierten Lesediagnostik die Verwendung der Begriffe wie folgt:

> „Der Begriff der Verlaufsdiagnostik wird in diesem Beitrag synonym mit Begriffen wie Lernfortschrittsdiagnostik (Walter, 2010), Lernverlaufsdiagnostik (Klauer, 2011), curriculumbasiertes Messen (CBM; Deno, 1985; Fuchs, 2004), systematische formative Evaluation (Fuchs & Fuchs, 1986) oder prozessorientierte Diagnostik (Souvignier & Förster, 2011) verwendet" (Walter, 2014, S.165).

4.1.2 Ziele und Methoden

> „Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited" (Black & Wiliam, 2009, S.9).

Wie die internationalen und nationalen Schulleistungstests wie TIMMS, IGLU und PISA gezeigt haben, liefern diese auf Statusdiagnostik ausgerichteten klassischen und normbasierten Tests wertvolle Informationen über die Kompetenzstände der Schüler*innen in den jeweiligen Domänen wie z.B. Lesen oder Schreiben. Solche Tests können aber „nicht hinreichend kurzfristige Leistungsveränderungen abbilden, so dass mit hoher Wahrscheinlichkeit kleinere oder nach kurzer Zeit erzielte Lernfortschritte, und damit wichtige Verbesserungen bei einzelnen Schülern, unentdeckt bleiben" (Walter, 2014, S.166). Der formativen Evaluation oder Diagnostik wird ferner das Potenzial zugeschrieben, Lernen fördern zu können, allerdings nicht unter allen Bedingungen wie Klauer (Klauer, 2014, S.2) pointiert hervorhebt: „Tatsächlich gibt es gut begründete Nachweise; dass die regelmäßige Leistungsmessung mit entsprechenden Rückmeldungen die Leistung deutlich verbessern kann, nur ist dies eben nicht unter allen Bedingungen der Fall." Eine alternative Bezeichnung für formative Leistungsevaluation ist der Begriff Lernverlaufsdiagnostik.

Lernverlaufsdiagnostik zielt darauf ab, nicht nur den Prozess des Lernfortschritts zu dokumentieren, sondern diesen durch Rückmeldung der Ergebnisse an Schüler*innen und Lehrkräfte zu fördern (Klauer, 2011) und dadurch eine nötige „Veränderung des Verhältnisses zwischen Unterricht und Diagnostik voranzutreiben“ (Walter et al., 2018, S.12). Die Vorläufer heutiger Lernverlaufsdiagnostiken reichen in die 1970er Jahre in den USA zurück (Deno & Mirkin, 1977, Deno, 1985; Deno, 2003; Fuchs & Fuchs, 1986). Tests zur Lernverlaufsdiagnostik messen die Veränderung eines Merkmales über einer bestimmte Zeitspanne, womit individuelle Lernverläufe sichtbar gemacht werden können (Wilbert & Linnemann, 2011). Dies erfordert Testinstrumente, die Lernen durch Verhaltensänderung mittels aufeinanderfolgender Messungen auf gleichem Anforderungsniveau in einem definierten Zeitintervall in einer unveränderten Situation beschreiben. Unter der Annahme, dass zwischen den Messungen eine Intervention (z.B. Rechtschreibunterricht) stattgefunden hat (Wilbert & Linnemann, 2011). An den Verhaltensveränderungen der Individuen kann „abgelesen“ werden, inwiefern es von einer spezifischen Intervention profitiert und ob es einen Lernzuwachs in einem bestimmten Aufgabenbereich gegeben hat.

Formative Leistungsdiagnostik kann auch innerhalb des Response-to-Intervention-Ansatz (RTI) verortet werden, indem Lernverlaufsdiagnostik als ein integraler Bestandteil des Unterrichts angesehen wird (Walter et al., 2018). Deno und Mirkin (1977) verfolgten mit ihrem Interventionsprogramm Data-Based Program Modification (DBPM) das Ziel, Tests zu entwickeln, anhand dessen der Lernverlauf von Kindern mit Lern- und Verhaltensproblemen in unterschiedlichen Domänen festgestellt und individuelle Förderpläne erstellt werden konnten. Die regelmäßige Leistungserhebung erfolgte auf eine sehr ökonomische Weise, indem die Ergebnisse computerbasiert ausgewertet und die Ergebnisse grafisch dargestellt wurden (Blumenthal et al., 2021). Mit dem sogenannten Curriculum-Based Measurement (CBM), ursprünglich für den sonderpädagogischen Kontext entwickelt, sollte es Lehrkräften ermöglicht werden, Lernverläufe von Schüler*innen nachvollziehen und Entscheidungen für notwendige Maßnahmen zu ihrer Unterstützung und Förderung treffen zu können:

> „The primary goal oft he research program was to develop measurement and evaluation procedures that teachers could use routinely to make decisions whether and when to modify a student‘s instructional program“ (Deno, 1985, S.221).

Curriculumbasierte Messverfahren (CBM) sind standardisierte, kurze und in regelmäßigen Zeitabschnitten eingesetzte Tests, die längsschnittlich Veränderungen und Entwicklungen im Lernverlauf von Kindern abbilden und eine frühzeitige Prävention durch passgenaue Förderung ermöglichen. Sie stellen eine Methode der Lernverlaufsdiagnostik dar und lassen auch Aussagen über die Wirksamkeit aktuell erfolgender Interventionen und Instruktionen zu (Voß & Hartke, 2014).

Walter und Clausen-Suhr (2018, S.12) fassen die Merkmale des Konzepts des CBM wie folgt zusammen:

> „Diese neue Klasse von Diagnostikverfahren ist allgemein dadurch gekennzeichnet, dass sie die klassischen Testgütekriterien wie Objektivität, Reliabi-

lität und Validität erfüllt, darüber hinaus das Gütekriterium der Änderungssensibilität mitbringen muss, schnell und unkompliziert anzuwenden ist, im Gegensatz zu den üblichen Schulleistungstests engmaschig (d.h. mehrfach im Schuljahr) angewandt werden kann, weil genügend Paralleltests zur Verfügung stehen, kurzfristige Veränderungen der Leistungspotenziale bei Schülern sensibel erfasst und damit präventiv möglichen Lernstörungen entgegenwirkt."

4.1.3 Gelingensbedingungen

Forschungsergebnisse zeigen, dass mit dem Einsatz von Instrumenten zur Lernverlaufsdiagnostik eine Vielzahl an positiven Effekten wie z.B. ein höherer Lernzuwachs (Förster & Souvignier, 2015; Förster & Souvignier, 2014) und eine bessere Unterrichtsqualität (Fuchs & Fuchs, 1986) verbunden ist. Kinder, die Lernprobleme bzw. einen sonderpädagogischen Förderbedarf haben, profitieren dabei im besonderen Maße von der Lernverlaufsdiagnostik (Stecker et al., 2005).

Inwieweit die positiven Effekte der Lernverlaufsdiagnostik zum Tragen kommen, hängt jedoch in hohem Maße auch vom Kenntnisstand der Lehrkräfte (Heward, 2003; Hintz & Grünke, 2009) und deren professionellem Umgang mit den aus diagnostischen Verfahren gewonnenen Informationen ab (Capizzi & Fuchs, 2005). Studienergebnisse zur Akzeptanz des Einsatzes von Instrumenten zur Lernverlaufsdiagnostik in den Domänen Lesen und Mathematik weisen darauf hin, dass Lehrkräfte den Einsatz von Instrumenten zur Lernverlaufsdiagnostik als Mehrwert ansehen, jedoch die kontinuierliche Erfassung des Lernfortschrittes der einzelnen Schüler*innen auch als einen Mehraufwand im Unterricht empfinden (Shapiro et al., 2005; Souvignier et al., 2014). Obwohl die diagnostische Tätigkeit und die Begleitung der Lernprozesse der Schüler*innen zu den Schlüsselkompetenzen einer Lehrkraft zählt (Kultusminister Konferenz, 2005), verfügen sie nicht immer im ausreichendem Maße über fachliche und diagnostische Kompetenzen im Bereich des Schriftspracherwerbs, um den Lernprozess ihrer Schüler*innen adäquat begleiten zu können und um deren Lern- und Leistungsstände richtig einschätzen zu können (Corvacho del Toro & Günther, 2013; Schröder, 2019). Schüler*innen, die von Lehrkräften mit einer hohen diagnostischen Kompetenz und einer hohen Fachkompetenz im Bereich Orthografie unterrichtet werden, haben im Vergleich zu anderen Schüler*innen eine höhere Rechtschreibkompetenz (Roos & Schöler, 2009). Damit steht und fällt der erfolgreiche Erwerb der Schriftsprache auch mit der jeweiligen Kompetenz der Lehrkraft (Roos & Schöler, 2009). Insbesondere beim Schriftspracherwerb ist jene Einschätzung jedoch elementar (Roos & Schöler, 2009), da die Lernausgangslagen der Schüler*innen in diesem Bereich sehr heterogen sind (Hanke & Schwippert, 2005; Weinhold et al., 2020). Darüber hinaus ist eine weitere Gelingensbedingung für formative Diagnostik, dass Lehrkräfte passende Förderimplikationen mit den Tests erhalten, um auf die Lernschwierigkeiten der Schüler*innen adäquat reagieren zu können (Stecker et al., 2005).

4.2 Instrumente zur Lernverlaufsdiagnostik im Bereich Rechtschreibung

Bei der anspruchsvollen Aufgabe, die individuellen Lernvoraussetzungen zu bestimmen, kann eine computergestützte Diagnostik eine hilfreiche Unterstützung für Lehrkräfte sein. Diese ist sehr ökonomisch bezüglich der Durchführung und Auswertung, ermöglicht eine differenziertere Aufgabenstellung (z.B. adaptives Testen) und stellt automatisiert differenzierte Feedbackinformationen zur Verfügung (Maier et al., 2016; Maier, 2014). Computerbasierte diagnostische Tests ermöglichen, dass die Schüler*innen die Tests, die standardisierte Testinstruktionen enthalten, unabhängig von der Lehrkraft selbstständig in ihrem individuellen Tempo bearbeiten können und eine automatisierte Auswertung und Ergebnisdarstellung möglich ist (Gebhardt et al., 2016).

In der psychologischen und sozialwissenschaftlichen Forschung gehört die computergestützte Messung zur Feststellung von Individual- und Gruppenunterschieden zum methodischen Standard (Goldhammer & Kröhne, 2020). Im Rahmen von PISA erfolgen die Schulleistungsvergleichsstudien z.B. in den Domänen Lesen, Mathematik und Naturwissenschaften seit 2015 ausschließlich computerbasiert. Aus diagnostischer Perspektive ergeben sich für eine computerbasierte Messung im Vergleich zur Paper-Pencil-Testung besondere Vorteile. Es können Items automatisiert generiert und im Sinne eines adaptiven Testens individuelle Aufgaben für unterschiedliche Schüler*innen zusammengestellt, automatisch ausgewertet und der Testablauf auf die jeweilige Leistung abgestimmt werden. Durch eine computerbasierte Auswertung der Testergebnisse kann direkt nach der Durchführung des Tests ein Feedback über die erbrachte Leistung ermöglicht werden, das sich positiv auf den Lernprozess auswirken kann (Goldhammer & Kröhne, 2020).

> „Das computerbasierte Assessment stellt damit ein Sammeln von empirischen Informationen unter Zuhilfenahme eines Computers im weiteren Sinne dar, wobei die Bedingungen für das Sammeln so gestaltet werden, dass auf Grundlage der gesammelten Informationen Schlussfolgerungen über Individual- und Gruppenunterschiede möglich sind. Der Computer wird dazu eingesetzt, die Items zu präsentieren (z. B. Text, Bild, Audio, Video, Simulation), ihre Abfolge zu steuern sowie die Interaktionen der Testperson mit der Aufgabe zu registrieren (z. B. über Mausklicks, Tastatur- und Touchdisplayeingaben) und ggf. automatisch auszuwerten“ (Goldhammer & Kröhne, 2020, S.121).

Bisher gibt es im deutschsprachigen Raum kein computerbasiertes Instrumenten zur Lernverlaufsdiagnostik für den Bereich Rechtschreibung für die Primarstufe, das über die Eingabe der Testergebnisse hinaus eine Durchführung bzw. Erhebung am PC erlaubt und von Schüler*innen durch eine intuitive Benutzeroberfläche eigenständig absolviert werden kann.

Während der Einsatz von Instrumenten zur Lernverlaufsdiagnostik allgemein, aber auch speziell für den Bereich der Rechtschreibung im angloamerikanischen Raum bereits auf eine lange Tradition zurückblicken kann, steht die Entwicklung solcher Verfahren für den

deutschsprachigen Raum noch am Anfang.Für den Grundschulbereich stehen zur formativen Erfassung der Rechtschreibleistung derzeit die beiden computerbasierten Programme Lernfortschrittsdiagnostik Orthographie (LDO) von Walter und Clausen-Suhr (2018) und die Lernfortschrittsdiagnostik RESI 1-4 (Blumenthal et al., 2020) zur Verfügung. Inwiefern sich die Verfahren hinsichtlich der Vorgehensweise bei der Testkonstruktion, den Auswertungskategorien und dessen theoretischen Grundlegungen unterscheiden, wird im Folgenden gezeigt. Die Instrumente werden unter dem Aspekt der zugrunde liegenden Theorien zum Schriftspracherwerb, zum Aufbau und zur Operationalisierung der Testitems analysiert sowie die Frage nach der Evidenz geklärt.

4.2.1 Lernfortschrittsdiagnostik Orthographie (LDO)

Die Lernfortschrittsdiagnostik Orthographie (LDO) von Walter und Clausen-Suhr (2018) ist ein Verfahren zur Beobachtung des Lernfortschrittes der orthografischen Kompetenz im Verlauf. Der Test kann zu fünf Messzeitpunkten eingesetzt werden und ist für die zweite und dritte Klasse konzipiert. Im LDO erfolgt eine Paper-Pencil Testung mittels Wortdiktaten, bei der die Kinder mit einem audiogestützten Programm zur Schreibung der Wörter angeleitet und die Ergebnisse in einen Computer gestützten Auswertungsassistenten von der Lehrkraft eingegeben werden können (Mau et al., 2018). Insgesamt sind 10 Diktate mit jeweils 23 Items im Testpool enthalten, die in zwei Varianten von fünf Diktaten pro Schuljahr eingesetzt werden können (Walter et al., 2018). Die LDO eignet sich sowohl als Einzel- als auch Gruppentest und die Bearbeitungszeit umfasst ca. 15 Minuten. Orthografische Kompetenz wird im LDO im Sinne der Ausführungen von Scheerer-Neumann (Scheerer-Neumann, 2007) und May (2010) verstanden (Walter et al., 2018). Im Zwei-Komponenten-Modell des Rechtschreibprozesses nach Scheerer-Neumann (Scheerer-Neumann, 2007, S.542) wird von zwei Gedächtniskomponenten (Gedächtnis: Regeln und Strukturen; Gedächtnis: Lexikon) ausgegangen, die beim Rechtschreibprozess eine Rolle spielen. Zum einen kann beim Schreiben auf einen sogenannten „Regelspeicher“ zurückgegriffen werden, wobei die Schreibung memoriert und aus dem innersprachlichen Gedächtnis abgerufen wird und zum anderen kann die Schreibung mit Hilfe eines mentalen Wort- bzw. Morphemlexikon konstruiert werden kann (Fay & Berkling, 2013; Scheerer-Neumann, 2007). Das Entwicklungsmodell des Rechtschreibkönnens nach May (2010) beschreibt den Prozess des Rechtschreibens anhand von Rechtschreibstrategien. Die Wortauswahl der LDO basiert auf zwei Wörterbüchern für die Grundschule (Walter et al., 2018). Die Operationalisierung der Items erfolgt anhand von sechs Kriterien, die als „Konstruktionsgerüst“ konzipiert wurden, um Paralleltests zu entwickeln (Walter et al., 2018, S.24-25):

- Kriterium 1: Die Anlautstruktur deutscher Wörter
- Kriterium 2: Die Wortarten und Wort-Endungen
- Kriterium 3: Das Graphem nach der Anlautstruktur
- Kriterium 4: Die Wortlänge

- Kriterium 5: Die Pluralformen
- Kriterium 6: Der Ausschluss von Wörtern

Die Auswertung erfolgt auf Ebene des Ganzwortes und auf Basis der Anzahl der Graphemtreffer in Anlehnung an May (2010). Zudem wird als Maß für die Lernfortschritte im Verlauf

> „der so genannte Orthografische Kompetenz-Index (OKI) eingeführt, der sich formelmäßig wie folgt darstellen lässt: OKI = GT x p(GW).
> Der OKI-Wert eines Schülers in einem Diktat setzt sich also aus der mit der Lösungswahrscheinlichkeit für ein Ganzwort (p(GW)) gewichteten Anzahl seiner Graphem-Treffer (GT) zusammen“ (Walter et al., 2018, S.15).

Dieser kann mit dem Auswertungs-Tool computergestützt berechnet werden. Auswertungsbereiche sind dabei wie folgt definiert (Walter et al., 2018, S.29):

- Orthographischer Kompetenz-Index (OKI):
  OKI = GT x p(GW)
- Ganzwort (GW):
  falsch geschrieben = 0 Punkte
  richtig geschrieben = 1 Punkte
- Graphem-Treffer (GT): Punkte entsprechend der Anzahl der richtig geschriebenen Grapheme
- Anlautstruktur (Anl):
  falsch geschrieben = 0 Punkte
  richtig geschrieben = 1 Punkte
- Groß- und Kleinschreibung (GK):
  falsch geschrieben = 0 Punkte
  richtig geschrieben = 1 Punkte

Die LDO wurde längsschnittlich an einer Einstichprobe (N = 1266) zu fünf Messzeitpunkten in der zweiten und dritten Klasse in sieben verschiedenen Bundesländern normiert. Statistische Analysen weisen zufriedenstellende bis gute Reliabilitäten des Instruments zwischen .76 und .97 auf. Es lassen sich statistisch signifikante Kompetenzentwicklungen im Bereich Rechtschreibung für die Klassenstufen zwei und drei mit dem LDO ermitteln (Walter et al., 2018). Die Ausführlichen Analysen sind bei Walter und Clausen-Suhr (2018) dargestellt.

4.2.2 Lernfortschrittsdiagnostik RESI 1-4

Die Lernfortschrittsdiagnostik RESI 1-4 (Blumenthal et al., 2020; Blumenthal et al., 2021) ist ein curriculumbasiertes Messverfahren zur Erfassung der Entwicklung der Rechtschreibkompetenzen, das in ein multimodales Diagnose- und Förderkonzept eingebunden ist. Die Lernfortschrittsdiagnostik besteht aus unterschiedlich langen Wortdiktaten für

den Primarbereich von der ersten bis zur vierten Klassenstufe. Das Verfahren ist im Gruppenverband durchführbar, die Bearbeitungszeit beträgt maximal 10 Minuten. Der Itempool besteht aus 407 Wörtern, die in einer Vorstudie pilotiert wurden (Voß et al., 2017). Insgesamt sind pro Klassenstufe zehn Parallel-Tests vorhanden, die in einem vierwöchigen Abstand eingesetzt werden können. In den einzelnen CBM sind zu einem Drittel Anker-Items (identische Wörter) unterschiedlicher Schwierigkeitsstufen enthalten, um Lernverläufe modellieren zu können (Blumenthal et al., 2020).

Das theoretische Konstrukt zur Operationalisierung und Strukturierung der Test-Items bildet das Kompetenzprofil Rechtschreibung nach (Reber & Kirch, 2013). „RESI 1–4 folgt einem Stufenaufbau (Reber & Kirch, 2013), in dem die regelhaft erworbenen Kompetenzen in den früheren Klassenstufen und die mit vielen Ausnahmen behafteten Kompetenzen den höheren Jahrgängen zugeordnet sind“ (Voß et al., 2020, S.93). In diesem Kompetenzprofil werden die Prinzipien der deutschen Orthografie (alphabetisch, phonolgisch, morphologisch, orthografisch, grammatisch) gemäß der Vorgehensweise vieler Lehrpläne in die drei Kompetenzstufen Mitsprechwörter, Nachdenkwörter und Merkwörter unterteilt und durch insgesamt 40 Rechtschreibstrategien operationalisiert (Reber & Kirch, 2013; Reber, 2017).

- In der Kategorie Mitsprechwörter sind jene Wörter mit regelhaften Phonem-Graphem-Korrespondenzen zusammengefasst „die Kinder nach der Strategie „Schreibe wie du sprichst“ ohne Kenntnis weiterer Regeln komplett richtig verschriftlichen“ (Reber, 2017, S.71).
- Die Kompetenzstufe Nachdenkwörter umfasst jene Wörter, die regelbasiert sind und nicht über eine Phonem-Graphem-Korrespondenz hergeleitet werden können.
- Jene Wörter, die in die Kategorie der Merkwörter fallen, stellen nach Reber und Kirch (2017) die schwierigste Stufe dar. Merkwörter können nicht regelbasiert hergeleitet werden und müssen auswendig gelernt werden.

Die Wörter werden den Schüler*innen von der Lehrkraft diktiert. Die Ergebnisse der Paper-Pencil Testung können anschließend auf der Onlineplattform www.lernlinien.de eingetragen und automatisiert ausgewertet werden. Es erfolgt eine quantitative Auswertung und Interpretation der Schreibungen im Vergleich zur Klasse und zur Altersnorm anhand von Prozenträngen (Blumenthal et al., 2020), wobei die Ergebnisse in Anlehnung an das Kompetenzprofil Rechtschreibung von Reber und Kirch (2013) zu Profilen der alphabetischen, phonologischen, morphologischen, orthografischen und grammatischen Strategie zusammengefasst werden (Blumenthal et al., 2020). Die Schreibungen werden auf der Grundlage von Graphemtreffern beurteilt, jedes richtig geschriebene Graphem wird mit einem Punkt bewertet. Wird z.B. das Wort <Schwein> orthografisch korrekt verschriftlicht, können vier Punkte erreicht werden „(z. B. Sch/w/ei/n = vier Grapheme = vier Punkte/Graphemtreffer)“ (Blumenthal et al., 2020, S.24).

Anhand des Kompetenzprofils nach Reber (2013) werden zudem Fehlerschwerpunkte identifiziert, wobei die jeweiligen Rechtschreibphänomene zu Niveaustufen zugeordnet und nach einem Ampelsystem bewertet werden (Blumenthal et al., 2020). Die Niveaustufen

geben Auskunft darüber, wie viel Prozent der im jeweiligen Wortmaterial repräsentierten Rechtschreibphänomene korrekt verschriftlicht wurden (Blumenthal et al., 2020).

- Niveau 1 : grün : 80-100%
- Niveau 2: gelb: 60-80%
- Niveau 3: rot: <60%

Dies stellt den Ausgangspunkt für die weitere differenzierte Diagnostik sowie für Übungen und Förderung dar (Blumenthal et al., 2020). Aus den Fehlerschwerpunkten lassen sich Informationen für die Förderung ableiten und es werden entsprechend Fördermaterialien zur Verfügung gestellt, die sich auf das Erlernen und Automatisieren von Rechtschreibregeln (regelbasierter Unterricht) beziehen:

- „Erarbeiten/ Wiederholen der notwendigen Vorläuferfähigkeiten zur Erarbeitung der Rechtschreibregel
- Einführen/ Erarbeiten der Rechtschreibregel
- Festigen der Anwendung der Rechtschreibregel
- Automatisierung der Anwendung der Rechtschreibregel inkl. Übungen zur Überprüfung auf Übergeneralisierung des Einsatzes der Regel“ (Blumenthal et al., 2020, S.29).

Es wird von einem stufenförmigen Schriftspracherwerb ausgegangen:

> „Bevor es an die Erarbeitung der jeweiligen Rechtschreibregel geht, wird das schriftsprachliche Vorwissen gesichert, um anschließend stufenweise die Kompetenzen der alphabetischen, phonologischen, morphologischen und orthografischen Strategie zu entwickeln“ (Blumenthal et al., 2020, S.29).

Die Lernfortschrittsdiagnostik RESI 1-4 wurde im Rahmen einer Studie zu fünf Messzeitpunkte im zweiten Halbjahr der ersten Klasse und ab dem ersten Halbjahr der zweiten Klasse bis zur vierten Klasse zu zehn Messzeitpunkten im Schuljahr 2016/2017 in Schulklassen (N = 28) in Mecklenburg-Vorpommern evaluiert. Zusammenfassend konnte für das Instrument eine zufriedenstellende Reliabilität zwischen .74 bis .89 ermittelt werden. Die meisten Items weisen eine ausreichende Stichprobenunabhängigkeit für die Teilungskriterien Median und Geschlecht und eine ausreichende Passung zum Raschmodell auf. Zudem bildet das CBM statistisch signifikante Kompetenzunterschiede über die Klassenstufen ab. Die Studie ist jedoch nicht repräsentativ, da die Evaluierung des Instruments nur in einem Bundesland (Mecklenburg-Vorpommern) erfolgte (Blumenthal et al., 2021). Die Ergebnisse sind ausführlich in Blumenthal et al. (2021) beschrieben.

Zusammenfassung

Die theoretische Grundlage eines diagnostischen Instruments hat einen großen Einfluss auf die Förderung und den Unterricht, da sie die Grundlage für die Entwicklung von Unterrichts- und Fördermaterialien sowie Lernaktivitäten bildet. Sie ermöglicht es den Lehrkräften, die Lernbedürfnisse und Fähigkeiten der Schüler*innen besser zu verstehen und darauf abgestimmte Förder- und Unterrichtsstrategien zu entwickeln. Die analysierten Verfahren unterscheiden sich hinsichtlich der Vorgehensweise bei der Testkonstruktion, den Auswertungskategorien und dessen theoretischen Rahmenkonzeptionen mit unterschiedlichen Annahmen über den Entwicklungs- und Erwerbsprozess des Schriftspracherwerbs, die wiederum mit unterschiedlichen didaktischen Konsequenzen für den Unterricht einhergehen. Gemein ist den beiden Testverfahren, dass sich die jeweils theoretischen Grundannahmen zum Schriftspracherwerb und zur Modellierung von Rechtschreibkompetenz einem normorientierten und statischen Kompetenzmodell (Dependenzhypothese) zuordnen lassen und die Kompetenzmodelle nicht validiert sind. Der Diskussionsstand zeigt, dass die hier verwendeten normbasierten Stufenmodelle des Schriftspracherwerbs, in denen das lautorientierte Schreiben als notwendiger Entwicklungsschritt auf dem Weg zum orthografischen Schreiben angesehen wird, den heterogenen Lernausgangslagen und Lernverläufen nicht gerecht werden können. Beide Testverfahren basieren in der quantitativen Auswertung auf der Zählung der richtig geschriebenen Grapheme (Graphemtreffer). Inwieweit es ein Regelwerk für eine differenzierte automatisierte Testergebnisanalyse vorliegt bzw. wie der Auswertungsalgorithmus funktioniert, wird nicht beschrieben.

4.3 Theoretische Grundlagen zur Konstruktion von Instrumenten zur Lernverlaufsdiagnostik

> „Psychologische Tests und Fragebögen haben das Ziel, Merkmalsträger (Testpersonen) hinsichtlich ihrer Merkmalsausprägungen einer metrisch vergleichenden Beurteilung zugänglich zu machen. Vor und während der Konstruktion eines Tests sind zahlreiche Aspekte zu berücksichtigen, die es erlauben, die Merkmalsausprägung zu quantifizieren und jeder Person einen (numerischen) Testwert zuzuordnen. Zur Beurteilung, ob eine solche Zuordnung von Testwerten zu Personen angemessen ist, werden verschiedene testtheoretisch basierte psychometrisch-statistische Maße herangezogen“ (Brandt & Moosbrugger, 2020, S.41).

Dieser Abschnitt bietet einen Überblick über den aufwendigen Prozess der Testentwicklung, der sich auf der Ebene einer Planungsphase und einer Konstruktionsphase vollzieht (Brandt & Moosbrugger, 2020).

4.3.1 Testplanung

In der Planungsphase geht es zunächst darum, das zu erfassende Merkmal einzugrenzen, zu definieren und den Geltungsbereich sowie die Zielgruppe des Tests zu bestimmen. Zudem müssen Entscheidungen bezüglich des Testumfangs, der Testlänge, der Testadministration sowie zum Testaufbau getroffen werden (Brandt & Moosbrugger, 2020).

Merkmalsdefinition

Zu Beginn steht die Präzisierung und Definition des Merkmals der Rechtschreibkompetenz im Fokus. Dies erfordert eine Literaturrecherche (Domänenanalyse), wobei vorhandene Theorien und empirische Befunde berücksichtigt werden sollen (Brandt & Moosbrugger, 2020, S.41):

„Bei einem **Merkmal** handelt es sich um eine (numerisch erfassbare) Variable, hinsichtlich derer sich verschiedene Personen (allgemeiner: Merkmalsträger) unterscheiden. Mit **Merkmalsausprägung** bezeichnet man eine quantitative oder qualitative Angabe darüber, welche Größe das Merkmal bei einer untersuchten Person aufweist."

Die Ausgangslage zur Merkmalsdefinition ist die diagnostische Fragestellung, welche die Ziele des Tests definiert. Die diagnostische Fragestellung dieser Arbeit lautet:

> „Wie entwickelt sich die Rechtschreibkompetenz von Grundschüler*innen im Lernverlauf auf Ebene der korrekt geschriebener Wörter und auf Ebene der orthographischen Teilkompetenzen?"

Als Grundlage für einen Test muss Rechtschreibkompetenz anhand von fachwissenschaftlichen Theorien begründet und anhand von validierten Modellen definiert werden. Es gilt auf Basis einer elaborierten Theorie ein Instrument zu entwickeln, das im Einklang mit konkreten Lehr- und Lernzielen steht. Im Rahmen einer Domänenanalyse müssen dazu verschiedene Sichtweisen auf das Konstrukt Orthografie analysiert und die unterschiedlichen Vorstellungen und Positionen diskutiert werden (vgl. Kap. 2.1.4), um eine Entscheidung treffen zu können, welche der Annahmen der unterschiedlichen Positionen zugrunde gelegt werden. Die Auswahl eines empirisch fundierten Rechtschreibkompetenzmodells ist wichtig, weil die Konsequenzen für den Unterricht und den anschließenden Implikationen, die aus jenen Modellen abgeleitet werden „in einem hohen Maß von den sprach- und lerntheoretischen Vorannahmen abhängig" sind (Hinney, 2010, S.50) und entscheidende Auswirkungen auf die Unterrichtsgestaltung haben können (Naujokat, 2015). Im Bereich Orthografie stellt dies eine besondere Herausforderung dar. Bevor ausgehend von der diagnostischen Fragestellung und den getroffenen Entscheidungen der Domänenanalyse und Domänenmodellierung der Test entwickelt werden kann, gilt es den Geltungsbereich und die Zielgruppe des Tests festzulegen (Brandt & Moosbrugger, 2020).

Geltungsbereich und Zielgruppe

Der Geltungsbereich eines Tests beschreibt die Einsatz- und Anwendungsmöglichkeiten sowie Aspekte der Validität (Brandt & Moosbrugger, 2020). Wichtig ist, dass mit dem Test zuverlässige Aussagen zur Beantwortung der diagnostischen Fragestellung getroffen werden können. Dazu müssen Kriterien zur Messgenauigkeit festgelegt werden:

> „...ein Verfahren, das für eine Individualdiagnostik verwendet werden soll (z. B. Test zur Diagnose von Rechenschwäche – TEDI-Math; Kaufmann et al. 2009), muss wesentlich genauer messen, als ein Screeningverfahren, das bei einer breiten Anwendung nur eine erste grobe Beurteilung des interessierenden Merkmals erbringen soll (z. B. das Depressions-Screening PRIME-MD; Spitzer et al. 1999)“ (Brandt & Moosbrugger, 2020, S.50).

Daneben spielt die Definition der Zielgruppe, über die Aussagen mit dem Test getroffen werden sollen, eine wichtige Rolle. Die Testaufgaben sollen auch zielgruppengerecht formuliert und sprachlich verständlich sein (Brandt & Moosbrugger, 2020).

Testlänge und Testzeit

Die Testlänge beschreibt die Anzahl der im Test enthaltenden Items (Lienert und Raatz, 1998). Die Testzeit gibt die Zeitspanne für die Bearbeitung der Testaufgaben vor. Die Testlänge ist von der Anzahl und Definitionsbreite der Konstrukte sowie von der Zielgruppe abhängig. Bei der Konzipierung einer differenzierten Individualdiagnose muss auf eine ausreichende Anzahl von Items des zu interessierenden Merkmals geachtet werden. Je mehr Items zur Erfassung eines Merkmalsbereichs im Itempool enthalten sind, desto präziser wird das Testergebnis.

> „Die für eine bestimmte Fragestellung notwendige Testlänge hängt auch von der Qualität der einzelnen Items ab. Die Konstruktion und die Auswahl/Selektion „guter“ Items ist somit von zentraler Bedeutung, um den Test möglichst reliabel und valide, aber gleichzeitig auch hinreichend kurz und ökonomisch zu gestalten“ (Brandt & Moosbrugger, 2020, S.52).

Testadministration

Im Rahmen der Testadministration werden grundsätzliche Entscheidungen über den Testmodus (papier- oder computerbasiert) sowie über die Form der Testung (Einzel- oder Gruppentestung) getroffen (Brandt & Moosbrugger, 2020). Unter Berücksichtigung des Aspekts der Zeitökonomie ist die Konzipierung eines computerbasierten Testdesign erforderlich. Damit entfällt die nachträgliche manuelle Eingabe der Testergebnisse durch die Lehrkraft. Zudem wird dadurch eine von der Testleitung unabhängige hohe Durchführungs- und Auswertungsobjektivität gewährleistet und eine ökonomische Testauswertung möglich, da die Antworten direkt erfasst werden können (Brandt & Moosbrugger, 2020). Durch eine automatisierte Testauswertung, die im Test implementiert ist, eröffnen sich große

Potenziale. Testreports und Rückmeldungen über die Testergebnisse können sofort nach Beendigung des Tests bereit gestellt werden (Goldhammer & Kröhne, 2020). Ein adaptives Testen wird ebenfalls dadurch möglich. Die Durchführung einer Individualdiagnostik erfordert häufig eine Einzeltestung, die mit einem erheblichen Aufwand im Vergleich zu einer wesentlich ökonomischeren Gruppentestung verbunden ist (Brandt & Moosbrugger, 2020).

Testaufbau

Der Aufbau eines Tests besteht typischerweise aus einer Testinstruktion, der konkreten Aufgabenstellung und der Erfassung demografischer Informationen (Brandt & Moosbrugger, 2020). Die Funktion der Testinstruktionen besteht darin, die Testpersonen zur Teilnahme zu motivieren und Informationen über die zu bearbeitenden Aufgaben zu liefern. Die Instruktionen sollen klare Erklärungen beinhalten, aus denen klare Handlungsweisen zur Beantwortung der Testaufgaben hervorgehen (Brandt & Moosbrugger, 2020). Dazu wird in der Regel ein Beispiel für eine Testaufgabe mit der dazugehörigen Antwort dem eigentlichen Test vorangestellt. Die konkreten Aufgabenstellungen müssen für jedes zu erfassende Merkmal und jede Merkmalsfacetten ausreichend Items enthalten. Die Testanweisungen und die Testmaterialien sollen sprachlich und optisch zielgruppengerecht sein, eine leichte Testbearbeitung ermöglichen und motivierend sein (Brandt & Moosbrugger, 2020).

4.3.2 Testkonstruktion

Die Konstruktion eines Tests, der einsetzbar ist und der den gängigen statistischen Gütekriterien entspricht, ist mit vielfältigen und aufwendigen Entwicklungsschritten verbunden (Brandt & Moosbrugger, 2020, S.58):

1. „Konstruktion/Generierung einer geeigneten Menge von Testaufgaben/Items („Itempool), einschließlich der Instruktion und der Wahl eines geeigneten Antwortformats
2. Qualitative Verständlichkeitsanalyse der Instruktion und der Items mit erforderlichen Nachbesserungen (erste Revision)
3. Erste empirische Erprobung der vorläufigen Testfassung („Pilotstudie“) an einer kleineren Stichprobe mit Itemanalyse und -selektion (zweite Revision)
4. Zweite empirische Erprobung („Evaluationsstudie“) an einer größeren, repräsentativen Stichprobe („Analysestichprobe“) mit psychometrischen Analysen und anschließender dritter Revision (ggf. sind weitere Revisionsschleifen erforderlich)
5. Abschließende Normierung der endgültigen Testform“

1. Operationalisierung der Testaufgaben

Die Konzeption und Operationalisierung von Testitems, deren theoretische Grundlegung sowie die Festlegung von Auswertungskategorien, anhand derer differenziert Aussagen über die Kompetenzentwicklung im Bereich Rechtschreibung getroffen werden können, ist die aufwendigste und bedeutendste Phase der Testentwicklung (Brandt & Moosbrugger, 2020).

Es werden Kriterien benötigt, anhand derer die Anforderungsstrukturen von Wörtern bzw. Rechtschreibkompetenz identifiziert und als Grundlage für die Erstellung von parallelen Testversionen genutzt werden können (Voß et al., 2017; Walter et al., 2018). Kompetenzmodelle eignen sich insbesondere als Grundlage für die Operationalisierung von Testaufgaben und zur Definition von Anforderungsniveaus.

> „Entscheidend ist die Forderung, die jeweilige Kompetenz, um die es im Einzelnen geht, durch eine Menge von Aufgaben zu definieren, zu deren Lösung die Kompetenz qualifiziert. Die Aufgabenmengen, deren Lösung beherrscht werden soll, sind streng im mengentheoretischen Sinne so zu definieren, dass für jede beliebige Aufgabe eindeutig entscheidbar ist, ob sie Element der Menge ist oder nicht“ (Klauer, 2011, S.213).

2. Verständlichkeitsanalyse

Eine erste qualitative Beurteilung des entwickelten Tests prüft die Verständlichkeit der Testaufgaben.

Ziel dabei ist die Identifikation inhaltlicher, praktischer und technischer Schwierigkeiten (Brandt & Moosbrugger, 2020). Die Erprobung sollte möglichst unter realistischen Bedingungen in der Praxis mit Testpersonen der Zielgruppe durchgeführt werden (Brandt & Moosbrugger, 2020). Bei sorgfältiger Itemkonstruktion ist eine Erprobung mit einer kleineren Stichprobe ausreichend (Brandt & Moosbrugger, 2020).

> „Die qualitative Überprüfung der Items soll in einer ersten Revision des Itempools münden, bei der alle Items mit Verständnisschwierigkeiten ausgesondert oder nachgebessert werden. Wird diese erste Revision nicht sorgfältig durchgeführt, so resultieren Mängel in der Testkonstruktion, die sich zu einem späteren Zeitpunkt auch nicht mit ausgefeilten statistischen Analysetechniken beheben lassen“ (Brandt & Moosbrugger, 2020, S.60).

3. Pilotstudie

Nachdem die Testkonstruktion und der vorläufige Itempool im Kontext der qualitativen Verständlichkeitsprüfung angepasst wurde, kann die vorläufige Testversion einer ersten empirischen Untersuchug (Pilotstudie) unterzogen werden, um die Qualität der konstruierten Items mittels deskreptiver Statistiken zu überprüfen.

4. Evaluationsstudie

Ziel der Evaluationsstudie ist es, statistische Berechnungen durchzuführen, um Erkenntnisse zur Passung mit dem zugrundegelegten psychometrischen Modell zu gewinnen (Brandt & Moosbrugger, 2020, S.63).

5. Testnormierung

Die letzte Phase der Testentwicklung besteht in der abschließenden Normierung der endgültigen Testform an einer großen, repräsentativen Stichprobe. Anhand der Testwerte der Zielgruppe können dann Normtabellen erstellt werden (Brandt & Moosbrugger, 2020).

4.4 Gütekriterien bei der Testkonstruktion zur kompetenzbasierten Lernverlaufsdiagnostik

Mit der Entwicklung von Instrumenten zur Lernverlaufsdiagnostik sind besondere Herausforderungen in Bezug auf die Operationalisierung der Items einschließlich deren theoretischen Fundierung und Zusammenstellung einer Aufgabenstichprobe mit gleichen Schwierigkeitsgrad, der Ergebnisauswertung sowie Evaluation verbunden (Fuchs, 2004; Klauer, 2011; Strathmann & Klauer, 2008; Strathmann et al., 2010; Voß et al., 2017; Walter et al., 2018; Wilbert & Linnemann, 2011). Ein Instrument zur Lernverlaufsdiagnostik muss hoch reliabel und zugleich veränderungssensitiv sein. Damit sich die Veränderungen in den Testergebnissen zwischen zwei Messzeitpunkten auf einen Lernzuwachs zurückführen lassen können, ist es erforderlich, dass die Aufgabenstichproben in ihrem Anforderungsniveau gleich sind und die Ergebnisse der einzelnen Items sich auf den gleichen latenten Faktor zurückführen lassen.

Als ein in der Forschungsgemeinschaft anerkannter Orientierungsrahmen haben sich u.a. die „Forschungsstufen“ nach Fuchs (2004, S.189):

- Technical features of the static score
- Technical features of slope,
- Instructional utility

sowie die Kriterien zur Analyse von Tests zur Lernverlaufsdiagnostik von Wilbert und Linnemann (2011) etabliert:

- Itemanalysen
- Dimensionalitätsprüfung
- Raschanalysen
- Testfairness

Die erste und zweite Forschungsstufe nach Fuchs (2004) beziehen sich auf die klassischen Gütekriterien Objektivität, Reliabilität, Validität sowie um die Sensibilität des Instruments und sind identisch mit den von Wilbert und Linnemann (2011) vorgeschlagen Standardanalysen zur Prüfung der Gütekriterien eines Tests. Wobei Wilbert und Linnemann (2011) auch den Aspekt der Testfairness in den Blick nehmen.

Forschungsstufen nach Fuchs

Fuchs (2004) nimmt in der Stufe Instructional utility die Aspekte der Praktikabilität und Effektivität eines CBM im Unterricht zusätzlich in den Fokus. Danach gilt es zu klären, wie es den Lehrkräften gelingen kann, anhand der durch den Einsatz von CBM gewonnenen Informationen, den Unterricht sowie die individuelle Förderung anzupassen (Deno, 2003) und den Lernzuwachs zu quantifizieren (Voß & Hartke, 2014). Die beschriebenen Anforderungen unterscheiden sich zwar nicht grundsätzlich von denen an Tests, mit denen eine Statusdiagnostik durchgeführt werden kann, bringen aber besondere Probleme bei der Umsetzung mit sich (Wilbert & Linnemann, 2011). Da es im Rahmen dieser Arbeit um die Entwicklung und Pilotierung CBM für den Bereich Rechtschreibung geht, werden im Folgenden vor allem die mit der Entwicklung eines Tests zur Lernverlaufsdiagnostik verbundenen hohen testtheoretischen Anforderungen bezüglich der Reliabilität, Validität, Eindimensionalität, der Vergleichbarkeit der einzelnen Tests sowie der Testfairness beschrieben (Wilbert & Linnemann, 2011).

Kriterien nach Wilbert

Ein Instrument zu Lernverlaufsdiagnostik muss hoch reliabel und zugleich veränderungssensitiv sein. Im Rahmen der klassischen Testtheorie (KTT) liegt eine hohe Reliabilität vor, wenn die Merkmalvarianz deutlich größer als die Fehlervarianz ist. Bei der Bestimmung der Retest-Reliabilität zur Überprüfung der Messgenauigkeit eines Testinstruments, ist es wichtig, dass der Messfehler in allen Skalenbereichen gleich ist (Wilbert & Linnemann, 2011). Voraussetzung für eine zuverlässige Einschätzung der Differenzwerte ist eine mindestens intervall skalierte Skala (Wilbert & Linnemann, 2011). Dies bedeutet, dass

> „bei Skalen, die sich durch die Aufsummierung einzelner Itemwerte ergeben, dass diese Items einen homogenen, unkorrelierten Messfehler aufweisen und in gleichem Maß von der zugrundeliegenden Merkmalsausprägung [...] beeinflusst sind“ (Wilbert & Linnemann, 2011, S.227).

Erfüllt ein Instrument zur Messung von Veränderung diese Anforderungen nicht, ist dieses ungeeignet.

Aus messtheoretischer Sicht zur Konstruktion eines Tests zur Lernverlaufsdiagnostik stellt die Rasch-Modellierung eine Lösung für die beschriebene Problematik dar, da die einzelnen beobachteten Werte eines Items in eine mindestens intervallskalierte Logitskala transformiert werden können (Wilbert & Linnemann, 2011). Anhand dessen lassen sich die Ausprägungen der Personenparameter innerhalb der Antwortkategorien eines Items

und zwischen den Antwortkategorien verschiedener Items auf Intervallskalenniveau ableiten (Wilbert & Linnemann, 2011). Tests zur Lernverlaufsdiagnostik müssen in ihrem Anforderungsniveau gleich sein, damit sich die Veränderungen in den Testergebnissen zwischen zwei Messzeitpunkten auf einen Lernzuwachs zurückführen lassen, da die Konzipierung von Paralleltests mit einem sehr hohen Aufwand verbunden und im Rahmen der klassischen Testtheorie zur Bestimmung der exakten Testschwierigkeit nur durch eine Trennung von Personen- und Itemparameter durchführbar ist (Wilbert & Linnemann, 2011). Im Rahmen der Probabilistischen Testtheorie ist die Bestimmung von Itemparametern hingegen unabhängig von den Personenparametern möglich. Voraussetzung für ein Instrument, das die Veränderung eines bestimmten Kompetenzbereichs misst, ist, dass die Ergebnisse der einzelnen Items auf den gleichen latenten Faktor zurückzuführen sind, womit die lokale stochastische Unabhängigkeit der einzelnen Testitems verbunden ist. Die Wahrscheinlichkeit, ein anderes Testitem zu lösen, verändert sich also durch das richtig Beantworten eines Items nicht (Wilbert & Linnemann, 2011). Des Weiteren spielt die Testfairness im umfassenden Sinne eine wichtige Rolle. Die Ergebnisse des Test dürfen „zu keiner systematischen Benachteiligung bestimmter Personen aufgrund ihrer Zugehörigkeit zu ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppen“ führen (Moosbrugger & Kelava, 2020, S.25). Wilbert & Linnemann (2011) schlagen bezogen auf die vorangegangene Skizzierung der besonderen Herausforderungen bei der Konstruktion von Tests zur Lernverlaufsdiagnostik zusammenfassend folgende Standardanalysen zur Prüfung der Gütekriterien eines Tests zur Lernverlaufsdiagnostik vor:

1. Itemanalysen
2. Dimensionalitätsprüfung
3. Raschanalysen
4. Testfairness

4.4.1 Analysen auf Basis der klassischen Testtheorie

Im Folgenden werden die deskriptiven Analysen auf Basis der klassischen Test Theorie (KTT) gemäß der Vorgehensweise nach Wilbert und Linnemann (2011) beschrieben. Anhand der Werte zur Itemschwierigkeit, Trennschärfe, interne Konsistenz, Retestreliabilität und Eindimensionalität der Items lassen sich erste Schlussfolgerungen ziehen, inwiefern die Items des Tests zur Abbildung der verschiedenen Merkmalsausprägungen der Testpersonen passen (Kelava & Moosbrugger, 2020a). Die Qualität eines Tests bzw. der im Test enthaltenen Items lässt sich an bestimmten statistischen Werten beurteilen.

Itemschwierigkeit

Die Itemschwierigkeit im Kontext der klassischen Test Theorie unterscheidet sich von der in der Item-Response-Theorie (Kelava & Moosbrugger, 2020a). Während in der KTT davon ausgegangen wird, dass die numerische Höhe des Schwierigkeitsparameters die Leich-

tigkeit bzw. Schwierigkeit eines Items angibt, kennzeichnet der Schwierigkeitsparameter im Rahmen der IRT die tatsächliche Schwierigkeit des Items (Kelava & Moosbrugger, 2020a). Items, die im mittleren Schwierigkeitsbereich ($P_i = 50$) liegen, differenzieren am besten zwischen Personen mit höheren und Personen mit niedrigeren Merkmalsausprägungen (Kelava & Moosbrugger, 2020a). Um zwischen sehr und weniger leistungsstarken Personen zu differenzieren, sollten die Items im Bereich von $5 \leq P_i \leq 20$ bzw. $80 \leq P_i \leq 95$ liegen. Eine gleichmäßige Verteilung von Items im Schwierigkeitsbereich von $5 \leq P_i \leq 95$ erlaubt eine Differenzierung über das ganze Merkmalsspektrum (Kelava & Moosbrugger, 2020a).

Trennschärfe

Ziel ist es, dass anhand der Items zwischen Personen mit hohen Merkmalsausprägungen und Personen mit niedrigen Merkmalsausprägungen gut differenziert werden kann. Hohe positive Trennschärfen deuten darauf hin, dass die einzelnen Items gut differenzieren. Trennschärfen, die im Bereich von 0.4 bis 0.7 liegen, sind als gut zu bewerten (Kelava & Moosbrugger, 2020a).

Interne Konsistenz

Die interne Konsistenz eines Tests wird im Rahmen der KTT anhand von Kennwerten zu Cronbachs Alpha, zur Retest- und Parallel-Reliabilität beurteilt. Die Schätzung der Reliabilität mit Cronbachs Alpha basiert auf den empirischen Varianzen und Kovarianzen der Itemvariablen. Die Berechnung der Retest- und Parallel-Reliabilität basiert auf den empirischen Korrelationen zwischen den Testwerten paralleler Tests (Schermelleh-Engel & Gäde, 2020). Zu beachten ist, dass der Wert Cronbachs Alpha im Zusammenhang mit der Anzahl der Items steht. Je mehr Items im Test enthalten sind, desto höher wird der Wert der Reliabilität der Skala (Schermelleh-Engel & Gäde, 2020).

Dimensionalitätsprüfung

Zur genaueren Überprüfung der Eindimensionalität der Items eignen sich faktoranalytische Verfahren bzw. testtheoretisch begründete Modellierungen der IRT. Diese Verfahren führen im Vergleich zu der vorangegangenen Beschreibung der deskriptiven Itemanalysen zu wesentlich belastbareren Ergebnissen (Kelava & Moosbrugger, 2020b). Anhand einer konfirmatorischen Faktorenanalyse kann geprüft werden, inwiefern eine Passung zwischen einem theoriebasierten Modell und den empirischen Daten besteht und inwieweit sich die Anzahl der postulierten Faktoren mit der jeweiligen Zuordnung der Items aus dem theoretischen Modell in den Daten replizieren lässt (Gäde et al., 2020).

Verzahnung von Theorie und Empirie

Bei der Auswahl geeigneter Items spielt neben den beschriebenen Itemanalysen auch die Auseinandersetzung mit dem theoretisch zugrunde gelegten Modell eine wichtige Rolle.

> „Eine rein von Kennzahlen getriebene Itemselektion ohne theoretische Auseinandersetzung mit den Iteminhalten ist nicht zweckmäßig und auch nicht im Sinne der Psychometrie. Vielmehr ist die psychometrische Itemanalyse ein iterativer Prozess der gelingenden Auseinandersetzung/Verzahnung von Theorie und Empirie“ (Kelava & Moosbrugger, 2020a, S.157).

4.4.2 Analysen auf Basis der Item-Response-Theorie

Im Unterschied zur klassischen Testtheorie, die ein Testergebnis als unmittelbare messfehlerbehaftete Merkmalsausprägung versteht, stellt das Testergebnis im Rahmen der IRT lediglich einen Indikator latenter Merkmale bzw. Verhaltensdispositionen (Latent Traits) dar, der über die Ausprägungen der manifesten Merkmale geschätzt wird (Döring & Bortz, 2016). Hinter der Item-Response-Theorie (IRT) steht die testtheoretische Grundannahme, dass das Antwortverhalten einer Person auf ein Item sowohl von der Eigenschaft des Items als auch von der Merkmalsausprägung der Person abhängt (Kelava & Moosbrugger, 2020b). Das Antwortverhalten einer Person auf ein Item ist also immer nur ein Indiz für das zugrunde liegende latente Konstrukt wie z.B. Rechtschreibkompetenz (Kelava & Moosbrugger, 2020a). Die IRT wurde von Thurstone (1928) entwickelt, erlangte aber erst durch die Entwicklung des Rasch-Modells (Rasch, 1960) und des Birnbaum-Modells (Birnbaum, 1968) an Bedeutung. Ziel ist es, Rückschlüsse auf die individuellen Ausprägungen der latenten Konstrukte ziehen zu können (Kelava & Moosbrugger, 2020b). Unterschieden wird innerhalb der IRT zwischen einer Vielzahl von unterschiedlichen statistischen, messtheoretischen und psychologischen Modellen (z.B. Rasch-Modell, Birnbaum-Modell etc.), die sich hinsichtlich der Anzahl der Antwortkategorien und zu schätzenden Itemparameter unterscheiden (Döring & Bortz, 2016; Kelava & Moosbrugger, 2020b).

Raschanalysen

Das Rasch-Modell ist das am weitesten verbreitete IRT-Modell. Das Birnbaum-Modell stellt eine Erweiterung des Rasch-Modells um einen itemspezifischen Trennschärfeparameter dar (Kelava & Moosbrugger, 2020b). Das Rasch-Modell wird den Latent-Trait-Modellen zugeordnet und findet eine sehr häufige Verwendung in der Leistungsdiagnostik, deren Ziel es ist, anhand von Testergebnissen auf eine bestimmte Kompetenz (latentes Merkmal) zu schließen (Kelava & Moosbrugger, 2020b). Latent-Trait-Modelle gehen im Gegensatz zu Latent-Class Modellen von einer Kontinuität des latenten Merkmals der Personenvariable aus. Die Antwort auf ein Testitem einer Person kann also auf ein kontinuierliches latentes Markmal (Trait) zurückgeführt werden, wobei sich die Ausprägung des latenten Merkmals zwischen den Personen unterscheidet (Kelava & Moosbrugger, 2020b).

Itemhomogenität

Im Rahmen der Rasch-Modellierung wird von einer Rasch-Homogenität der Items ausgegangen, die eine spezielle Form der Eindimensionalität darstellt.Obwohl die Items je nach Itemschwierigkeit unterschiedliche Anforderungsniveaus aufweisen, messen sie dasselbe latente Merkmal (Kelava & Moosbrugger, 2020b).

> „Unter der Rasch-Homogenität versteht man, dass den Antworten auf alle Items eines Tests genau *e*ine latente Variable n (nämlich das interessierende Merkmal) zugrunde liegt und dass – abgesehen von den variierenden Itemschwierigkeiten β_i– genau diese eine latente Personenvariable die Unterschiede im Antwortverhalten der verschiedenen Personen erzeugt (und in gewisser Weise auch erklärt)“ (Kelava & Moosbrugger, 2020b, S. 373).

Mittels einer Itemcharakteristischen Funktion (IC-Funktion) kann der Zusammenhang zwischen der Lösungswahrscheinlichkeit eines Items und dem zu untersuchenden latenten Merkmal hergestellt werden (Kelava & Moosbrugger, 2020b).

Item- und Personenparameter

Eine weitere Voraussetzung für die Parameterschätzungen und Modellkontrollen ist die lokale stochastische Unabhängigkeit der Antworten von Personen auf die Testitems (Kelava & Moosbrugger, 2020b). Item- und Personenparameter müssen geschätzt werden, da

> „das Rasch-Modell die Wahrscheinlichkeit einer Datenmatrix und der in ihr enthaltenen Antwortmuster in Abhängigkeit von Item- und Personenparametern beschreiben“ (Kelava & Moosbrugger, 2020b, S. 387).

IRT-Modelle mit dichotomen Antwortformaten stellen die einfachste Ausgangssituation dar, um zu überprüfen, inwieweit Rückschlüsse auf das dem Modell zugrunde liegende latente Personenmerkmal gezogen werden können (Kelava & Moosbrugger, 2020b). Zur Schätzung von Item- und Personenparameter kommen in der IRT Maximum-Likelihood-Schätzmethoden (ML-Schätzverfahren) und Bayes‘sche Schätzverfahren zum Einsatz, die sich jeweils in weitere Verfahren mit verschiedenen Schätzalgorithmen mit unterschiedlichen Eigenschaften ausdifferenzieren lassen (Rose, 2020). Zu den wichtigsten ML-Verfahren zählen die Joint ML (JML), Conditional ML (CML) und Marginal ML (MML) (Rose, 2020). Die Schätzverfahren berücksichtigen auf unterschiedliche weise die Item- und Personenparameter. So erfolgt z.B. beim JML Verfahren u.a. die gemeinsame Schätzung der Item und Personenparameter, während beim CML Verfahren die Personenparameter nicht mitgeschätzt werden (Rose, 2020). Das Bayes‘sche Schätzverfahren stellt ein alternative Vorgehensweise bei der Schätzung von Item- und Personenparameter dar. Das CML-Verfahren wird im Rahmen von Raschmodellierungen häufig angewendet (Rose, 2020).

Schätzung der Modellparameter

Die auf der Basis der Likelihood-Schätzung getroffenen Modellannahmen sind noch nicht valide und erfordern in einem weiteren Schritt die Überprüfung der Modellkonformität. Dies kann an unterschiedlichen Parametern wie z.B. der Itemschwierigkeit oder Trennschärfe beurteilt werden (Kelava & Moosbrugger, 2020b). Infit-Statistiken geben darüber Auskunft, inwiefern die jeweiligen Items personenübergreifend zum Messmodell passen. Viele Raschanalyseprogramme berechnen hierzu den gewichteten Mean-Square der Infit- und Outfit Werte, dessen Erwartungswert 1 ist (Bond et al., 2020). Werte des gewichteten Mean-Squares im Bereich von $.75 \leq \text{Infit} \leq 1.3$ liegen im akzeptablen Bereich (Bond et al., 2020). Hohe positive Trennschärfen deuten darauf hin, dass die einzelnen Items gut differenzieren. Trennschärfen, die im Bereich von 0.4 bis 0.7 liegen, sind als gut zu bewerten (Kelava & Moosbrugger, 2020a).

Modellvergleich

Liegt keine Modellpassung vor, empfiehlt sich eine Itemselektion anhand der Item-Fit-Indizes oder eine Erweiterung der Modellannahmen zu Gunsten eines 2- bzw. 3-PL Modells. Sowohl der Trennschärfe Parameter als auch die Item-Fit-Indizes werden kontrovers diskutiert, da die Item-Fit-Indizes nicht stichprobenunabhängig sind und die Diskrimination eines Items als geeignetes Verfahren zum Ausschluss eines Items in Frage gestellt wird (Embretson, 1996). Während im Rahmen der KTT die Reliabilität der messfehlerbehafteten Testvariablen bestimmt wird, wird die Reliabilität der geschätzten latenten Personenwerte im Kontext der IRT modellbasiert bestimmt (Schermelleh-Engel & Gäde, 2020). Anhand dessen kann die Information abgelesen werden, inwiefern sich die Unterschiede zwischen den geschätzten Personenwerte auf tatsächliche Unterschiede zwischen den Personen beziehen. Die Interpretation der Werte erfolgt analog zur klassischen Testtheorie (Schermelleh-Engel & Gäde, 2020).

Testfairness

Testfairness liegt vor, „wenn die resultierenden Testwerte zu keiner systematischen Benachteiligung bestimmter Personen aufgrund ihrer Zugehörigkeit zu ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppen führen“ (Moosbrugger & Kelava, 2020, S.25). Personen mit gleichem Fähigkeitsniveau, die verschiedenen Gruppen angehören, erreichen also die gleichen Personenscores in einem Test (Koller et al., 2012). Die Überprüfung der Subgruppeninvarianz anhand von externen Teilungskriterien (Geschlecht) wird als Überprüfung von DIF bezeichnet (Koller et al., 2012). Ist ein Item für Personen mit gleicher Fähigkeit unterschiedlich schwer zu lösen, liegt Differential Item Functioning (DIF) vor. Dies kann darauf hinweisen, dass durch das Item eine weitere Eigenschaft, wie z.B. sprachliche Kompetenz gemessen wird und die Validität des Items nicht gegeben ist (Wilbert & Linnemann, 2011). Die Methode des DIF lässt auch Aussagen zur Konstruktvalidität eines Tests zu (Wilbert & Linnemann, 2011).

4.5 Zusammenfassung

Eine individuelle Lernverlaufsdiagnostik gilt als ein wichtiges Instrument zur Verbesserung von Bildungsqualität. Nicht immer verfügen Lehrkräfte im ausreichendem Maße über fachliche und diagnostische Kompetenzen im Bereich des Schriftspracherwerbs, um den Lernprozess ihrer Schüler*innen adäquat zu begleiten und um deren Lern- und Leistungsstände richtig einschätzen zu können (Corvacho del Toro & Günther, 2013; Schröder, 2019).

Der Bedarf an geeigneten, praktikablen und ökonomischen Instrumenten zur Lernverlaufsdiagnostik für die Praxis ist im Bereich Orthografie nach wie vor aktuell. Für die Grundschule liegen für diesen Bereich derzeit die zwei computergestützten Programme Lernfortschrittsdiagnostik Orthographie (LDO) von Walter et al. (2018) und die Lernfortschrittsdiagnostik RESI 1-4 (Blumenthal et al., 2020, Blumenthal et al., 2021) vor. Die Verfahren unterscheiden sich hinsichtlich der Vorgehensweise bei der Testkonstruktion, den Auswertungskategorien und deren theoretischen Rahmenkonzeptionen mit unterschiedlichen Annahmen über den Entwicklungs- und Erwerbsprozess des Schriftspracherwerbs, die wiederum mit unterschiedlichen didaktischen Konsequenzen für den Unterricht einhergehen. Gemein ist den beiden Testverfahren, dass sich die jeweils theoretischen Grundannahmen zum Schriftspracherwerb und zur Modellierung von Rechtschreibkompetenz zu einem normorientierten und statischen Kompetenzmodell (Dependenzhypothese) zuordnen lassen und die Kompetenzmodelle nicht validiert sind. Die Annahme eines stufenförmigen Entwicklungsprozesses der normorientierten Kompetenzmodelle gilt als eine starke Abstraktion des tatsächlichen Schriftspracherwerbs und ist durch aktuelle Forschungsergebnisse widerlegt. Die Entwicklung von Instrumenten zur Lernverlaufsdiagnostik erfordert eine neue theoretische Fundierung, die auf den tatsächlichen Schriftentwicklungen und sprachsystematischen Erkenntnissen basiert. Es liegt kein Instrument zur Lernverlaufsdiagnostik für den Primarbereich vor, dessen theoretische Grundlegung auf einem sprachsystematischen Kompetenzmodell [1] basiert, die sprachlichen Voraussetzungen der Kinder als Lernausgangslage berücksichtigt und eine quantitative und qualitative Fehleranalyse auf Basis eines validierten Rechtschreibkompetenzmodells ermöglicht.

[1] Blatt und Frahm (2013, S.16) verstehen unter dem Begriff „Sprachsystematisch“, „dass die Schriftsprache aus graphematischer Forschungssicht zum einen eine eigenständige Systematik aufweist und zum anderen zusammen mit dem System der gesprochenen Sprache das umfassende Sprachsystem bildet.“

5 Ziele und Fragestellungen der Studie

Das Erlernen der Grundfertigkeiten des Schreibens gilt im Primarbereich als eine in den Bildungsstandards für den Deutschunterricht normativ gesetzte Kompetenzerwartung (Kultusminister Konferenz, 2005), die zum Ende der Grundschulzeit von allen Schüler*innen als Regelstandard erreicht werden soll. Die Fähigkeit, orthografisch normgerecht schreiben zu können, hat Einfluss auf die Lernerfolge in allen Schulfächern und ist für den weiteren Bildungsverlauf eines Kindes grundlegend.

Für eine frühzeitige und engmaschige Begleitung der individuellen Lernverläufe im Bereich Rechtschreibung sprechen die Studienergebnisse der empirischen Bildungsforschung zur Rechtschreibkompetenz deutscher Grundschüler*innen, sowie die Ergebnisse aus der Lehr-Lernforschung zum domänenspezifischen Wissen von Lehrkräften und deren diagnostischen Kompetenzen im Bereich Rechtschreibung. Bezug nehmend auf die heterogenen Lernausgangslagen und Lernprozesse beim Erwerb von Rechtschreibkompetenz ist die regelmäßige Bestimmung der Lernstände maßgebend. Das Konzept der Lernverlaufsdiagnostik gilt vor diesem Hintergrund als das wichtigste Instrument für erfolgreiche Lernprozesse und als ein wichtiges Instrument zur Verbesserung von Bildungsqualität. Nur so können präzise Einschätzungen bezüglich der individuellen Lernentwicklungen vorgenommen und Schwierigkeiten präventiv beim Lernen entgegengewirkt werden. Ergebnisse nationaler und internationaler Studien zeigen, dass mit dem Einsatz von Instrumenten zur Lernverlaufsdiagnostik eine Vielzahl an positiven Effekten verbunden sind, wie z.B. ein höherer Lernzuwachs und eine Verbesserung der Unterrichtsqualität. Auch diese Befunde sprechen für einen frühzeitigen und regelmäßigen Einsatz im Primarbereich. Entgegen der Bedeutung der individuellen Lernverlaufsdiagnostik für gelingende Lernprozesse im Bereich Rechtschreibung ist die Umsetzung des Konzepts im deutschsprachigen Raum nicht zufriedenstellend realisiert. Es fehlen geeignete Instrumente.

Bisher gibt es für den Primarbereich kein webbasiertes Instrument zur kompetenzbasierten Lernverlaufsdiagnostik auf Basis des sprachsystematischen Rechtschreibkompetenzmodell, das für die Schulpraxis über eine Onlinelernplattform zugänglich ist[1]. Außerdem fehlt die -wegen des Paradigmenwechsels innerhalb der Schriftlichkeitsforschung längst überfällige- Anwendung einer sprachsystematischen Modellierung von Rechtschreibkompetenz und infolgedessen einer sprachsystematischen Fundierung des Unterrichts zum Schriftspracherwerb. Daran anknüpfend stellt sich die Frage, wie ein Instrument zur Lernverlaufsdiagnostik im Bereich Rechtschreibung auf der Basis eines sprachsystematischen Rechtschreibkompetenzmodells konstruiert und anschließend für die Schulpraxis zugänglich gemacht werden kann.

[1] Das Rechtschreibkompetenz-Messverfahren (ReKoMe) ist seit 2017 auf der Onlinelernplattform www.Levumi.de bisher unter dem Namen „Wortdiktat“ (Mau, 2017) implementiert.

In Anlehnung daran müssen die mit der Lernverlaufsdiagnostik verbundenen neuen Anforderungen an die Lehrkräfte berücksichtigt werden. Eine ökonomische Durchführung und eine automatisierte Auswertung ist vor dem Hintergrund der Praktikabilität und der Akzeptanz für den Einsatz von Instrumenten zur Lernverlaufsdiagnostik im Schulalltag zentral. Ziel der Studie ist daher die Konstruktion, Implementation, Pilotierung und Evaluation eines webbasierten Rechtschreibkompetenz-Messverfahrens (ReKoMe) zur differenzierten Lernverlaufsdiagnostik für Schüler*innen der dritten Klasse, das ökonomisch im Unterricht eingesetzt werden kann und die sprachlichen Voraussetzungen und heterogenen Lernausgangslagen der Kinder durch die Verwendung eines adäquaten und validierten Rechtschreibkompetenzmodells sowie Ergebnisse der Sprach- und Schriftlichkeitsforschung berücksichtigt (Blatt et al., 2015; Hinney, 2010).

Vor dem Hintergrund fehlender webbasierter und praxiszugänglicher Testinstrumente zur Lernverlaufsdiagnostik im Bereich Rechtschreibung - insbesondere auf der Grundlage der sprachsystematischen Sichtweise auf den Schriftspracherwerb und den bisherigen Forschungsdesideraten - hat die Studie einen explorativen Charakter. Infolgedessen werden Forschungsfragen formuliert. Die zentrale Forschungsfrage der Arbeit lautet:

> **Forschungsfrage 1:** Inwiefern kann auf Basis des sprachsystematischen Rechtschreibkompetenzmodells ein webbasiertes, zeit- und testökonomisches Instrument zur differenzierten Lernverlaufsdiagnostik im Primarbereich entwickelt werden?

Daraus resultiert eine weitere Forschungsfrage zur Testkonstruktion:

> **Forschungsfrage 2:** Inwiefern kann ein Algorithmus entwickelt werden, der die Testergebnisse auf Ganzwortebene und auf Ebene orthografischer Teilkompetenzen automatisiert analysiert, codiert und zuverlässige Ergebnisse liefert?

Zwei weitere Forschungsfragen schließen sich an, die sich auf die empirische Überprüfung des Tests beziehen:

> **Forschungsfrage 3:** Ist das Rechtschreibkompetenz-Messverfahren (ReKoMe) ein reliables und valides Instrument zur differenzierten Lernverlaufsdiagnostik von Rechtschreibkompetenz in der dritten Grundschulklasse?
>
> **Forschungsfrage 4:** Lässt sich die theoretisch postulierte mehrfaktorielle Struktur des sprachsystematischen Rechtschreibkompetenzmodells für die dritte Klasse empirisch nachweisen?

Ziel der explorativen Studie ist es, ein webbasiertes Instrument zu konstruieren, dass Lehrkräfte dabei unterstützen kann, frühzeitig Probleme beim Schriftspracherwerb aufzudecken und passende Fördermaßnahmen einleiten zu können. Am Ende sollen die Lernentwicklungen der Schüler*innen differenziert und zuverlässig mit dem Test eingeschätzt und Aussagen darüber gemacht werden können, inwiefern sich der Rechtschreibunterricht auf die individuelle Kompetenzentwicklung der Kinder auswirkt. Ferner soll die Testauswertung und Darstellung der Ergebnisse in Lerngraphen und Balkendiagrammen durch eine automatisierte differenzierte Fehleranalyse erfolgen, die die Grundlage für die Konzipierung individueller Förder- und Lernangebote ist. Ein grundlegendes Verständnis der

Orthografiesystematik ist die Grundvoraussetzung, um einen an den individuellen Lernständen orientierten Rechtschreibunterricht realisieren zu können. Durch die Implementierung des Tests auf einer Onlinelernplattform ist ein leichter Zugang für die Schulpraxis möglich.

Zur Durchführung des Forschungsvorhabens werden die folgende Schritte vollzogen:

1. Konstruktion und Operationalisierung: Operationalisierung der Testaufgaben für das ReKoMe auf Basis des sprachsystematischen Rechtschreibkompetenzmodells sowie Festlegung von Auswertungskategorien. Als theoretisches Rahmenkonzept liegen im sprachsystematischen Rechtschreibkompetenzmodell fünf Prinzipien vor, denen jeweils Teilkompetenzen zugeordnet werden. Diese stellen unterschiedliche Anforderungen an die Schreibenden dar und bilden die Teilkompetenzen in der Rechtschreibung ab.

2. Digitale Umsetzung und Implementation: Entwicklung eines Algorithmus für eine automatisierte Ergebnisanalyse, Konzipierung eines webbasierten Testdesigns, Entwicklung einer Tastaturschulung, Implementierung des ReKoMe auf Onlinelernplattform Levumi.

3. Pilotierung: Durchführung einer Paper-Pencil Studie und einer Prototypenstudie, um die Qualität der konstruierten Testaufgaben, des Algorithmus und des webbasierten Testdesigns des ReKoMe unter realistischen Bedingungen im Schulsystem zu überprüfen. Die Ergebnisse dieser Studien dienen dazu, die Integrierbarkeit des ReKoMe im Unterricht zu bewerten und mögliche Verbesserungen zu identifizieren.

4. Evaluation: Überprüfung der psychometrischen Güte des konstruierten ReKoMe mittels Analysen der klassischen Testtheorie und der Item-Response-Theorie sowie die Überprüfung der faktoriellen Struktur des zugrunde gelegten theoretischen Rahmenmodells.

6 Konstruktion des ReKoMe

Dieses Kapitel beschreibt den Konstruktionsprozess des Rechtschreibkompetenz-Messverfahrens (ReKoMe) (vgl. Kap. 6.2), die digitale Umsetzung und Implementierung (vgl. Kap. 6.3) einschließlich der Entwicklung des Algorithmus für die automatisierte Ergebnisanalyse (vgl. Kap. 6.3.1), die Entwicklung einer Tastaturschulung sowie die zusammenfassende Darstellung der konstruierten Instrumente (vgl. Kap. 6.4). Wichtige methodische Aspekte der Testkonstruktion werden auf der Grundlage der bereits dargestellten theoretischen Überlegungen und Forschungsbefunde sowie der Definition der Testanforderungen in Kapitel 6.1 berücksichtigt. In Kapitel 6.2 stehen die Operationalisierung der Testaufgaben, ihre theoretische Grundlegung und die Festlegung von Auswertungskategorien im Fokus. Diese Kategorien ermöglichen es, differenzierte und theoretisch fundierte Aussagen über die Rechtschreibkompetenzentwicklung zu treffen. Der Algorithmus ist für die automatisierte quantitative und qualitative Analyse der Ergebnisse von zentraler Bedeutung. In Kapitel 6.4 werden schließlich die konstruierten Versionen des ReKoMe (vgl. Kap. 6.4.1 und vgl. Kap. 6.4.2) sowie die entwickelte Tastaturschulung (vgl. Kap. 6.4.3) dargestellt.

6.1 Methodisches Vorgehen

Die Konstruktion eines Tests zur Lernverlaufsdiagnostik, der einsetzbar ist und der den gängigen statistischen Gütekriterien entspricht, ist mit vielfältigen und aufwendigen Entwicklungsschritten verbunden (Brandt & Moosbrugger, 2020). In der Planungsphase für die Konstruktion des ReKoMe geht es zunächst darum,

1. das zu erfassende Merkmal durch eine theoriegeleitete Auswahl eines Kompetenzmodells einzugrenzen und zu definieren,
2. den Geltungsbereich und die Zielgruppe des ReKoMe festzulegen,
3. Entscheidungen über den Testmodus, die Implementation und den Testaufbau zu treffen,
4. die heterogenen Vorerfahrungen im Umgang mit einem PC durch die Entwicklung einer Tastaturschulung zu berücksichtigen,
5. Kriterien für die Testauswertung festzulegen,
6. ein geeignetes Regelwerk für die Entwicklung des Algorithmus auszuwählen.

Theoriegeleitete Auswahl eines Kompetenzmodells zur Merkmalsdefinition

Das diagnostische Ziel des zu konstruierenden ReKoMe ist die differenzierte Erfassung der individuellen Rechtschreibkompetenzentwicklung von Schüler*innen in der Primarstufe und bildet die Ausgangslage zur Merkmalsdefinition. Die Auswahl einer domänenspezifischen Theorie ist bei der Entwicklung eines Testinstruments zur kompetenzbasierten Leistungsmessung zentraler Standard der empirischen Bildungsforschung. Anhand von Kompetenzmodellen lässt sich der Anforderungsbereich der Testaufgaben strukturieren und zuordnen. Dieses Vorgehen macht erst eine objektive, reliable und valide Aussage über die Kompetenzen in der Zieldomäne möglich. Die Auswahl eines empirisch fundierten Rechtschreibkompetenzmodells ist wichtig, weil die Konsequenzen für den Unterricht und den anschließenden Implikationen, die aus jenen Modellen abgeleitet werden „in einem hohen Maß von den sprach- und lerntheoretischen Vorannahmen abhängig“ sind (Hinney, 2010, S.50) und entscheidende Auswirkungen auf die Unterrichtsgestaltung haben können (Naujokat, 2015). Dazu gilt es, vorhandene Theorien und empirische Befunde zur Domäne Rechtschreibung in den Blick zu nehmen, unterschiedliche Positionen zu analysieren und zu diskutieren, um eine Entscheidung darüber treffen zu können, welches Kompetenzmodell dem zu entwickelnden Test zugrunde gelegt werden soll (vgl. Kap. 2.1.4).

Im Bereich Orthografie stellt dies eine besondere Herausforderung dar. Obwohl zahlreiche Modelle zum Schriftspracherwerb vorliegen, die den idealtypischen Schriftspracherwerb abbilden, besteht kein Konsens darüber, wie die orthografische Kompetenz zu modellieren und zu operationalisieren ist. Vielmehr werden unterschiedliche Standpunkte vertreten und verschiedene Erwerbsmodelle als Ausgangspunkt für die Kompetenzmodellierung gewählt (Hinney, 1997; Naujokat, 2015) mit jeweils unterschiedlichen Schlussfolgerungen für die Schulpraxis (Reichardt, 2015). Zudem fehlt vielen Erwerbsmodellen eine ausreichende empirische Grundlage (Becker, 2008). Die bisherigen Ausführungen haben bereits gezeigt, dass das sprachsystematische Rechtschreibkompetenzmodell den kontextspezifischen Anforderungen von Rechtschreibkompetenz und den dazugehörigen domänenspezifischen Eigenschaften gerecht wird. Es eignet sich im besonderen Maße für die theoriegeleitete Aufgabenentwicklung, die kognitive Entwicklungsmerkmale und längsschnittliche Kompetenzentwicklungen berücksichtigen. Das sprachsystematische Kompetenzmodell ist daher die theoretische Grundlage zur Konstruktion des ReKoMe. Als theoretisches Rahmenkonzept für den Test liegen im sprachsystematischen Rechtschreibkompetenzmodell fünf Prinzipien vor, denen jeweils Teilkompetenzen zugeordnet werden, die unterschiedliche Anforderungen an die Schreibenden darstellen und Rechtschreibkompetenzen abbilden (vgl. Abb. 6.1).

Die Automatisierung der Teilkompetenzen stellt die Voraussetzung für die Weiterentwicklung der individuellen Rechtschreibkompetenz dar. Im sprachsystematischen Modell wird Rechtschreibkompetenz als ein komplexes kognitives Konstrukt erfasst, differenziell überprüft und als Ausdruck einer jeweils spezifischen Kompetenz im Sinne von Hartig und Klieme (2006) verstanden (Blatt et al., 2015). Die Rechtschreibfähigkeit vollzieht sich in Abhängigkeit der individuellen Lernvoraussetzungen mit der Zeit und integriert sich in

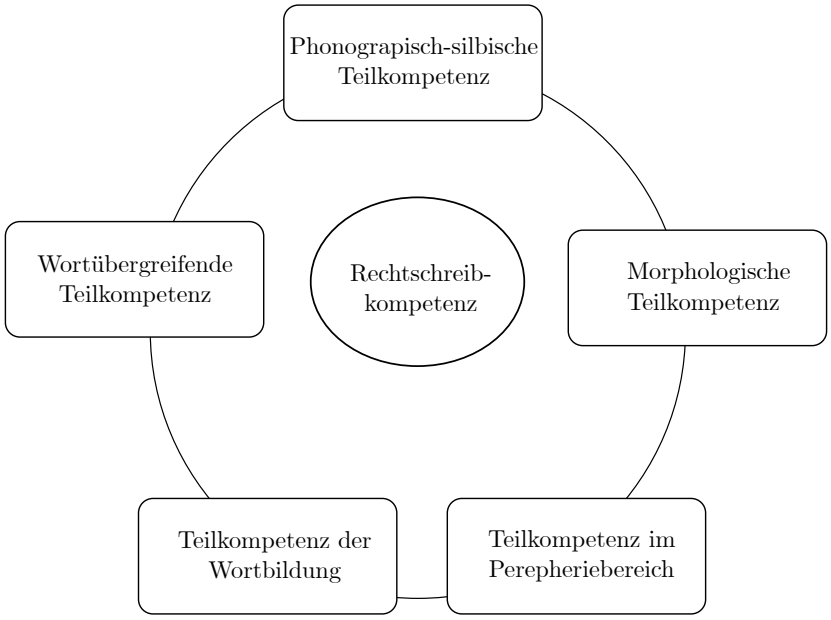

Abbildung 6.1: Rahmenkonzeption zum sprachsystematischen Rechtschreibtest (Blatt et al., 2015, S.237)

ein komplexes Modell der Rechtschreibkompetenz (Blatt & Frahm, 2013). Die Automatisierung der Teilkompetenzen ist die Voraussetzung für die Weiterentwicklung von Rechtschreibkompetenz. Im Unterschied zu anderen Entwicklungsmodellen geht man von keiner strikten Abfolge von bestimmten Phasen aus, sondern benennt einzelne Teilkompetenzen, die der Rechtschreibung zugeordnet werden können. Nachdem das Merkmal Rechtschreibkompetenz anhand einer ausgewählten fachwissenschaftlichen Theorie begründet und ein validiertes Rechtschreibkompetenzmodell ausgewählt wurde, muss der Geltungsbereich sowie die Zielgruppe des Tests festgelegt werden (Brandt & Moosbrugger, 2020).

Geltungsbereich und Zielgruppe

Der Geltungsbereich eines Tests beschreibt die Einsatz- und Anwendungsmöglichkeiten sowie Aspekte der Validität (Brandt & Moosbrugger, 2020). Wichtig ist, dass mit dem Test zuverlässige Aussagen zur Beantwortung der diagnostischen Fragestellung getroffen werden können. Dazu müssen Kriterien zur Messgenauigkeit festgelegt werden.

Das ReKoMe wird als Instrument zur Individualdiagnostik konstruiert und muss daher im Vergleich zu einem Screeningverfahren sehr messgenau sein (Brandt & Moosbrugger, 2020). Die Durchführung einer Individualdiagnostik erfordert häufig eine Einzeltestung, die mit einem erheblichen Aufwand im Vergleich zu einer wesentlich ökonomischeren Gruppentestung verbunden ist (Brandt & Moosbrugger, 2020). Das ReKoMe soll für Einzeltestungen und Gruppentestungen einsetzbar sein.

Die Zielgruppe des ReKoMe sind Schüler*innen der dritten Klasse. Der inhaltliche Schwerpunkt liegt vor allem auf den Teilkompetenzen des phonographisch-silbischen und dem morphologischen Prinzip des sprachsystematischen Rechtschreibkompetenzmodells. Häufig bestehen noch in der dritten Klasse Probleme bei der korrekten Anwendung dieser Prinzipien.

Webbasierter Testmodus und Implementation

Grundvoraussetzung für einen praktikablen Einsatz des ReKoMe im Unterricht ist, dass es eigenständig von Schüler*innen durchgeführt werden kann. Das ReKoMe wird als webbasiertes Verfahren konstruiert, wodurch sich wichtige zeit- und testökonomische Vorteile ergeben:

1. Eine automatisierte Antwortregistrierung und Testauswertung ermöglicht eine zeitökonomische Ergebnisanalyse, da die nachträgliche manuelle Eingabe und Auswertung entfällt.
2. Testreports und Rückmeldungen über die Testergebnisse können sofort nach Beendigung des Tests bereitgestellt werden (Goldhammer & Kröhne, 2020).
3. Ein adaptives Testen ist möglich.
4. Es ist eine von der Testleitung unabhängige hohe Durchführungs- und Auswertungsobjektivität gewährleistet (Brandt & Moosbrugger, 2020).

Für die Implementation von ReKoMe eignet sich insbesondere eine Onlinelernplattform. Onlinelernplattformen stellen einen komplexen technischen Rahmen für digitales Lernen zur Verfügung und bieten Möglichkeiten zur Darstellung und Analyse von formativen Testergebnissen (Maier, 2014). Sie bilden die klassische Hierarchiestruktur der Schule ab und sind eine ideale Möglichkeit, wichtige Funktionsbereiche für Blended-Learning zu implementieren (Maier, 2014).

Durch die Implementierung des ReKoMe auf der Onlinelernplattform Levumi ergeben sich im Vergleich zu einem zu installierenden Programm bzw. gegenüber einem auf einem Datenspeicher verfügbaren Programm viele Vorteile (Gebhardt et al., 2016):

- es gibt keine spezifischen technischen Voraussetzungen, die die Schulcomputer zusätzlich erfüllen müssen,
- die Daten können von unterschiedlichen Geräten (Handy, Tablet, PC) ortsungebunden abgerufen werden,
- die Datensicherung ist im Vergleich zu einer lokalen Sicherung auf einem PC höher,
- auftretende administrative Probleme bei der Bedienung des Programms können dezentral von den Administratoren der Internet-Seite behoben werden,
- Erweiterungen, Aktualisierungen und Verbesserungen am Test lassen sich leicht implementieren,

- es entstehen keine Kosten z.B. für einen technischen Support in der Schule.

Testaufbau

Der Aufbau eines Tests besteht meistens aus einer Testinstruktion, der konkreten Aufgabenstellung und der Erfassung demografischer Informationen (Brandt & Moosbrugger, 2020). Die Funktion der Testinstruktionen besteht darin, die Testpersonen zur Teilnahme zu motivieren und Informationen über die zu bearbeitenden Aufgaben zu liefern. Die Instruktionen sollen klare Erklärungen beinhalten, aus denen klare Handlungsweisen zur Beantwortung der Testaufgaben hervorgehen (Brandt & Moosbrugger, 2020). Zu diesem Zweck wird dem eigentlichen Test in der Regel eine Beispielaufgabe mit der entsprechenden Antwort vorangestellt. Die konkreten Aufgaben müssen für jedes zu erfassende Merkmal und jede Merkmalsfacette eine ausreichende Anzahl von Items enthalten. Die Testinstruktionen und Testmaterialien sollen sprachlich und visuell zielgruppengerecht gestaltet sein, eine leichte Testbearbeitung ermöglichen und motivierend wirken (Brandt & Moosbrugger, 2020).

Der Einsatz von Audiomaterial bietet ein Höchstmaß an Standardisierung, da die Instruktionen zur Testdurchführung immer gleich sind. Die Schüler*innen können zudem die Lautstärke, die Abspielzeitpunkte sowie die Abspielhäufigkeit individuell bestimmen (Goldhammer & Kröhne, 2020).

Tastaturschulung

Eine sichere und benutzerfreundliche Testumgebung ist wichtig, damit die Vorkenntnisse im Umgang mit Computern keinen behindernden Faktor darstellen (Parshall et al., 2010). Die Berücksichtigung der heterogenen Vorerfahrungen der Schüler*innen im Umgang mit einem Computer und einer Tastatur ist daher zentral, der Umgang mit dem jeweiligen Eingabegerät muss vor der Durchführung des ReKoMe eingeübt werden (Goldhammer & Kröhne, 2020). Für das ReKoMe wird eine Tastaturschulung entwickelt.

Kriterien zur Testauswertung

Im englischsprachigen Raum gilt die Auswertung der Schreibungen auf Ganzwortebene (quantitativ) als das Mittel der Wahl zur ökonomischen Erfassung der Rechtschreibleistung (Hosp & Hosp, 2003; Hosp et al., 2016). Dabei gehen jedoch wichtige Informationen für eine individuelle Förderung verloren (Fay et al., 2012). Für eine differenzierte Lernverlaufsdiagnostik ist eine automatisierte Ergebnisanalyse, die über das bloße Auszählen von richtig und falsch geschriebenen Wörtern hinausgeht und die Art der Schreiblösungen in den Blick nimmt, eine Grundvoraussetzung, da eine manuelle qualitative Fehleranalyse sehr zeitaufwändig ist (Frahm, 2013). Darüber hinaus zeigen die Ergebnisse qualitativer Rechtschreibfehleranalysen, dass Rechtschreibprobleme auch bei gleichen quantitativen Ausprägungen in klar abgrenzbaren Bereichen liegen können (Corvacho del Toro, 2016).

Für das ReKoMe werden Auswertungskategorien auf zwei Ebenen konstruiert. Es wird...

- auf Wortebene geprüft, wie viele Wörter insgesamt richtig und wie viele Wörter falsch geschrieben wurden.
- auf Ebene der Teilkompetenzen des sprachsystematischen Kompetenzmodells die Art der Schreiblösung innerhalb eines Wortes näher analysiert, um differenzierte Lernentwicklungsprofile erstellen zu können, anhand derer sich gezielte Förderimplikationen ableiten lassen.

Auf diese Weise ist eine Beschreibung der Testantworten in Bezug auf den Grad ihrer Systematik möglich.

Entwicklung eines Algorithmus

Die festgelegten Auswertungskategorien erfordern die Entwicklung eines neuen Algorithmus für eine automatische, differenzierte, quantitative und qualitative Testauswertung im Rahmen dieses Forschungsvorhabens. Eine manuelle Codierung einer qualitativen Ergebnisanalyse erfordert auch grundlegende Kenntnisse der Sprachwissenschaften, die bei Lehrkräften nicht vorausgesetzt werden können und aus ökonomischen Aspekten im Unterrichtsalltag nicht praktikabel sind (Frahm, 2013). Die automatische Testauswertung auf Wortebene ist leicht durch ein automatisiertes Abgleichen mit der korrekt hinterlegten Lösung im Algorithmus zu realisieren (Goldhammer & Kröhne, 2020). Im Hinblick auf eine qualitative Testauswertung, die nur eine bestimmte Buchstabenkombination bzw. Struktureinheit (Lupenstelle) innerhalb der Testantwort fokussiert, ist dies wesentlich komplexer. Bei der Analyse der Testergebnisse ist es notwendig, auch Struktureinheiten zu identifizieren, die trotz korrekter Schreibung der Lupenstelle als falsch gewertet werden, da vor oder nach der Lupenstelle falsche Buchstaben hinzugefügt werden können (Frahm, 2013).

Die Entwicklung eines Kategoriensystems zur differenziellen Fehleranalyse von Schülerschreibungen auf Basis eines Kompetenzmodells ist komplex und erfordert die Konzipierung eines Regelwerks, anhand dessen die Struktureinheiten und Ausschlüsse der Wörter definiert werden können (Frahm, 2013). Die besondere Herausforderung bei der Analyse einer bestimmten Lupenstelle auf Korrektheit ist, dass nicht nur die Lupenstelle an sich, sondern auch die sich im Umfeld befindlichen Buchstaben untersucht werden müssen (Frahm, 2013). Frahm (2013, S.151) hat im Rahmen der Entwicklung eines Programms zur computerbasierten Auswertung eines Rechtschreibtests für die fünfte Klasse auf Basis des sprachsystematischen Rechtschreibkompetenzmodells dafür ein Regelwerk erstellt:

1. „Ist der Vokal gesondert ausgewiesen (# ä, # äu - Morphologie), muss durch einen Ausschluss verhindert werden, dass eine fälschliche Dehnung oder Dopplung als richtig codiert wird: <Änderungsschneiderei>: # nder falsch, wenn davor <nn>: Ännderungsschneiderei

2. Bei Singularformen muss durch einen Ausschluss verhindert werden, dass die Pluralform als richtig codiert wird: <Videospiele>: # spiele falsch, wenn danach <n>: *Videospielen

3. Bei den Flexionsendungen # t und # d muss die Codierung von <td> als richtig mit einem Ausschluss verhindert werden: <entspannt>: # t falsch, wenn davor <d>: *entspanndt

4. Bei zusätzlichen Buchstaben (<e>, <s>) nach Wortstämmen müssen die Einheiten mit Hilfe von Ausschlüssen als falsch codiert werden: <Drehbrücke>: # Dreh falsch, wenn danach <e>: *Drehebrücke

5. Bei Doppelkonsonanten vor Suffixen, die mit Vokal beginnen (<ung>, <ei>,<ig>, <isch>, <innen>), muss die Codierung als richtig mit Ausschlüssen (<l>, <r>,<n>, <s> vehindert werden: <Nationalität>: # ität falsch, wenn davor <ll>: *Nationallität

6. Bei Präfixen, die auf einem Vokal enden, müssen <h> und <r> als Ausschluss eingefügt (<nach>, <zu>, <ge>) werden: <zuverlässig>: <zu> falsch, wenn danach <h>,<r>,<ff>, *zuhverlässig

7. Bei Präfixen (<ver>, <vor>, <un>), die auf einen Konsonanten enden, wird die Codierung von Doppelkonsonanten als richtig verhindert, indem ein Ausschluss eingefügt wird: # ver falsch, wenn danach <r>: *zuverrlässig“.

Anhand dessen können Struktureinheiten identifiziert werden, die trotz korrekter Schreibung der Lupenstelle als falsch gewertet werden müssen, da vor oder nach der Lupenstelle falsche Buchstaben hinzugefügt wurden. Dies lässt sich an dem Beispielwort <Änderungsschneiderei> verdeutlichen. In dem Wort <Änderungsschneiderei> soll die Lupenstelle #nder, die der Teilkompetenz des phonographischen-silbischen Prinzips zugeordnet wird, geprüft werden. Es liegt die folgende Fehlschreibung vor <Ännderungsschneiderei>, die analysiert werden soll. Obwohl die Struktureinheit #nder richtig verschriftlicht wurde, zeigt die Schreibung, dass die Einsicht in das Phonographisch-Silbische Prinzip noch fehlt, da der Konsonant verdoppelt wurde (Frahm, 2013). Nach dem Regelwerk von Frahm (2013) müssen folglich für diese Lupenstelle die weiteren Buchstaben <n> und <h> als Ausschlüsse gelten, sofern sie der Lupenstelle z.B. <Ähnderungsschneiderei> vorangestellt werden.

Frahm (2013) zeigt in ihrer Studie, dass sich auf der Grundlage des Regelwerks eine automatisierte Codierung entwickeln lässt, die im Vergleich zur manuellen Codierung zuverlässiger ist. Dies stellt die Grundlage für ein formatives Assessment auf Basis des sprachsystematischen Rechtschreibkompetenzmodells dar, da nur durch eine ökonomische und automatische Codierung eine sofortige Ergebnisauswertung möglich ist (Frahm, 2013).

Testanforderungen

Aus den bisherigen Ausführungen resultieren zusammenfassend folgende Testanforderungen:

- Der Aufbau und die Bedienung des ReKoMe soll im Verständnis und in der Anwendung kindgerecht und motivierend sein und eine leichte Navigation durch den Test erlauben. Wichtig ist, dass der Test eigenständig von den Schüler*innen durchgeführt werden kann. Es muss ein praktikables, diagnostisches Instrument für den Unterrichtsalltag sein.
- Das Antwortformat des ReKoMe wird offen gestaltet. Die Beantwortung der Testaufgaben erfolgt durch die Schüler*innen schriftlich per Tastatur. Dadurch entfällt eine aufwendige, fehlerbehaftete nachträgliche manuelle Eingabe der Testergebnisse.
- Die Wörter und die entscheidenden Buchstabenkombinationen der Testantworten werden für die Feststellung der Rechtschreibkompetenzentwicklung einschließlich der jeweiligen Teilkompetenz automatisch durch einen entwickelten Algorithmus codiert und anschließend differenziert analysiert.
- Die Testergebnisse müssen valide Informationen über Leistungsfortschritte, Leistungsrückschritte und Leistungsstagnation liefern. Die Auswertung muss mit einer manuellen fachlichen Codierung vergleichbar sein.
- Der Test muss für die Praxis differenzierte Lernentwicklungsprofile graphisch in Lerngraphen und -balken darstellen. Daraus müssen sich individuelle Förderimplikationen ableiten lassen.

6.2 Operationalisierung der Testaufgaben

Mit der Entwicklung von Instrumenten zur Lernverlaufsdiagnostik sind besondere Herausforderungen im Hinblick auf die Konzeption und Operationalisierung der Items, der Zusammenstellung einer Aufgabenstichprobe mit gleichem Schwierigkeitsgrad und der Testauswertung (Strathmann & Klauer, 2008; Strathmann et al., 2010; Voß et al., 2017; Walter et al., 2018) sowie im Hinblick auf die Evaluation verbunden (Fuchs, 2004; Klauer, 2011; Wilbert & Linnemann, 2011). Das Ziel des ReKoMe ist es, anhand der Testaufgaben zuverlässige, differenzierte Aussagen über die Rechtschreibkompetenzentwicklung von Schüler*innen im Primarbereich treffen zu können. Als theoretisches Rahmenkonzept zur Operationalisierung der Testitems liegen im sprachsystematischen Rechtschreibkompetenzmodell fünf Prinzipien vor (vgl. Kap. 3.2.2), denen jeweils Teilkompetenzen zugeordnet werden. Diese stellen unterschiedliche Anforderungen an den Schreibenden dar und bilden Rechtschreibkompetenz ab. Die Automatisierung der Teilkompetenzen stellt die Voraussetzung für die Weiterentwicklung von Rechtschreibkompetenz dar. In den Anforderungsbereichen des Kernbereichs fallen alle Rechtschreibphänomene, die regelbasiert herleitbar und durch einen Wissenstransfer erlernbar sind. Dazu zählen das phonographisch-silbische, morphologische, wortübergreifende und Wortbildungs Prinzip.

Der Peripheriebereich umfasst die Ausnahmen, die die Kinder sich überwiegend durch Üben einprägen müssen. Da es um die Konzipierung eines Tests zur differenzierten Individualdiagnose geht, ist auf eine ausreichende Anzahl von Items des zu interessierenden Merkmals im Itempool zu achten. Je mehr Items zur Erfassung eines Merkmalsbereichs im Itempool enthalten sind, desto präziser wird das Testergebnis. Die Verwendung des sprachsystematischen Rechtschreibkompetenzmodells lässt die Erfassung mehrerer Teilkompetenzen bzw. Merkmalsfacetten mit einer Testaufaufgabe zu und bietet so eine besonders ökonomische Testung. Zur Auswahl eines geeigneten Wortpools gibt es wenig Anhaltspunkte. In einer Analyse, inwiefern sich die Empfehlungen der Kultusministerkonferenz (Kultusminister Konferenz, 2005), die Rahmenrichtlinien für den Deutschunterricht (LISUM, 2004) und Grundwortschatzlisten einen Orientierungsrahmen zur Generierung eines Wortpools als hilfreich erweisen könnten, resümieren Voß, Sikora und Mahlau (2017, S.186) „dass es nur wenige Anhaltspunkte für die Definition eines Wortschatzkorpus´[sic] für den Bereich Rechtschreibung in der Primarstufe gibt“.
Für das ReKoMe werden deshalb auf der Grundlage des theoretischen Rahmenmodells 53 Testwörter zielgruppenadäquat ausgewählt und anhand des sprachsystematischen Rechtschreibkompetenzmodells operationalisiert. Der Itempool enthält eine ausreichende Menge an Testitems, damit die einzelnen Phänomene der Teilkompetenzen ausreichend repräsentiert sind. Jedes Item prüft eine oder mehrere Teilkompetenzen des sprachsystematischen Rechtschreibkompetenzmodells durch Lupenstellen. In der folgenden Tabelle wird exemplarisch aufgezeigt, wie die Wörter entsprechend der Teilkompetenzen in Struktureinheiten aufgegliedert sind.

Prinzip	Phonographisch-Silbisches Prinzip	Morphologisches Prinzip	Peripheriebereich	Prinzipien der Wortbildung
Korb		# Korb		
spülen	# sp; # len	# ü		
bissig	# biss			# ig
Sahne	# Sa; # ne		# h	
Meer		# Meer		
Video			# Video	

Tabelle 6.1: Zuordnung der Teilkompetenzen in Struktureinheiten

Die orthografisch korrekte Schreibung der Struktureinheiten #len, #Sa und #ne kann mit Hilfe von Phonem-Graphem-Korrespondenzen geschrieben werden, wenn die sogenannte Explizitlautung als phonologische Referenz dient. Zum phonographisch-silbischen Prinzip im Kernbereich gehören auch der Silbenanfangsrand #sp und die Silbengliederung in #biss. So muss z. B. im Wort <spülen> die silbenstrukturelle Information erkannt werden, dass die Struktureinheit #sp entgegen der Graphem-Phonem-Korrespondenz (GPK) nicht mit <schp>, sondern mit <sp> geschrieben wird, um eine Überlänge im Schriftbild zu vermeiden. Das einsilbige Wort #Korb und die Struktureinheiten #ü sind dem mor-

phologischen Prinzip im Kernbereich zugeordnet. Die Schreibweisen lassen sich durch die Rückführung auf die Stammschreibweise herleiten. Das Dehnungs-h in dem Wort #Sahne sowie die Vokalbuchstabenverdopplung #Meer als Längenmarkierung lassen sich nicht regelbasiert herleiten und gehören zum Randbereich der Rechtschreibung. Die Großbuchstaben #K, #S, #M und #V gehören zum wortübergreifenden Prinzip. Dieses Prinzip wird bei der Testauswertung der Ergebnisse der Schüler*innen zunächst nicht ausgewertet, da die Bedienung der Tastenkombination am PC für die Großschreibung für die Kinder schwer sein könnte und somit nicht die tatsächliche Kompetenz gemessen werden würde.

Im Fokus des ReKoMe stehen insbesondere das phonographisch-silbische und das morphologische Prinzip. Kinder mit Problemen beim Schriftspracherwerb haben häufig noch in der dritten Klasse Probleme bei der korrekten Verschriftlichung dieser Prinzipien. Im Itempool sind auch Testaufgaben zum Peripheriebereich und zum Prinzip der Wortbildung enthalten, damit alle Teilkompetenzen der Rechtschreibkompetenz erfasst werden können. Dies ist wichtig, wenn die Rechtschreibprobleme nicht im Bereich des phonologischen-silbischen oder dem morphologischen Prinzip liegen. Sonst würden Schwierigkeiten in anderen wichtigen Teilbereichen der globalen Rechtschreibkompetenz unerkannt bleiben. Die Lupenstellen, welche die jeweiligen Teilkompetenzen repräsentieren, sind im Itempool deshalb unterschiedlich gewichtet. Die Anzahl der Lupenstellen der jeweiligen Teilkompetenzen ist in Tabelle 6.2 dargestellt.

	Phonographisch-Silisches Prinzip	Morphologisches Prinzip	Peripheriebereich	Prinzipien der Wortbildung
Lup	29	24	13	5

Tabelle 6.2: Verteilung der Lupenstellen nach Teilkompetenzen des sprachsystematischen Rechtschreibkompetenzmodells

In Anlehnung an Strathmann et al. (2010) wurden in einem zweiten Schritt die Testwörter jeweils in einen Satzkontext als Hilfestellung für das Wortverständnis eingebettet wie z.B. für das Testwort Blume: „Die Blume ist gelb. Blume.“

6.3 Digitale Umsetzung und Implementation

Die digitale Umsetzung des Tests, die Entwicklung eines Algorithmus zur automatisierten Auswertung der Testitems und die Konzipierung sowie Implementierung des Tests auf der Onlinelernplattform www.levumi.de ist Gegenstand dieses Abschnittes.

Zur Entwicklung des Algorithmus und der digitalen Umsetzung des Tests entstand eine Kooperation mit Prof. Dr. Andreas Mühling vom Institut für Informatik an der Christian-Albrechts-Universität zu Kiel. Für die interdisziplinäre Zusammenarbeit war die Autorin am Institut für Informatik in der Abteilung „Didaktik der Informatik“ Prof. Dr. Andreas

Mühling als Gastwissenschaftlerin tätig und für alle Fachinhalte bei der technischen Umsetzung des Algorithmus verantwortlich. Im Rahmen einer Bachelorarbeit (Albert, 2017) wurde der Algorithmus für die automatisierte, quantitative und qualitative Ergebnisanalyse auf Basis des sprachsystematischen Rechtschreibkompetenzmodells programmiert und in einer weiteren Bachelorarbeit (Bastian, 2017) eine neue Struktur für die Implementierung des Tests auf der Onlineplattform „Levumi“ für den Bereich Rechtschreibung geschaffen.

6.3.1 Entwicklung des Algorithmus zur automatisierten Testauswertung

Eine manuelle Codierung einer qualitativen Fehleranalyse ist aus ökonomischen Aspekten nicht praktikabel und erfordert grundlegende Kenntnisse der Sprachwissenschaften, die bei Lehrkräften nicht vorausgesetzt werden kann (Frahm, 2013).

Durch eine automatisierte Testauswertung, die im Test implementiert ist, eröffnen sich große Potenziale. Testreports und Rückmeldungen über die Testergebnisse können sofort nach Beendigung des Tests bereitgestellt und entsprechende Hinweise zur Förderung gegeben werden (Goldhammer & Kröhne, 2020). Dies spielt im Rahmen des formativen Testens eine wichtige Rolle, da sich eine Rückmeldung zum Lernstand unmittelbar positiv auf die Lernentwicklung auswirken kann (Goldhammer & Kröhne, 2020). Ein adaptives Testen wird ebenfalls dadurch möglich. Dafür ist es notwendig, einen Algorithmus zu entwickeln.

Grundlage für die Entwicklung des Algorithmus ist das Regelwerk zur computerbasierten Codierung von Wörtern auf Basis des sprachsystematischen Rechtschreibkompetenzmodells von Frahm (2013) (vgl. Kap. 6.1). Der Algorithmus basiert auf der Levenshtein-Distanz (1966) und Levenshtein-Spur, die in der Praxis häufig zur Bestimmung der Ähnlichkeit von Zeichenketten, z. B. bei der Rechtschreibprüfung verwendet werden. Die Levenshtein-Distanz gibt die Anzahl der Aktionen von Einfüge-, Lösch- und Ersatzoperationen an, die zur Umwandlung von einer Zeichenkette in eine zweite durchgeführt werden muss und wird z.B zur Überprüfung der Rechtschreibung verwendet. Der Algorithmus basiert auf der Levenshtein-Distanz (1966) und der Levenshtein-Spur, die Basis für den Algorithmus bildet die Kategorisierung der Testitems in Lupenstellen, Wortstämme, Präfixe sowie Pluralformen. Alle Buchstaben der Testitems werden in die jeweilige Teilkompetenz des sprachsystematischen Rechtschreibkompetenzmodells kategorisiert. Anhand dieser Kategorisierung wird die Schreibung der Kinder mit den hinterlegten Informationen zu den orthorafisch korrekten Schreibungen der Testitems geprüft.

Dies soll am folgenden Beispiel für das Wort <Versteck> exemplarisch dargestellt werden.

Beispiel: Versteck ,3 -4 ,| ,5 -7 , false ,| ,0 -2 , true ,| ,| , Verstecke

In der ersten Spalte ist die orthografisch korrekte Schreibweise des Items hinterlegt. In der Spalte zwei bis fünf werden die jeweiligen Struktureinheiten den entsprechenden Teilkompetenzen zugeordnet. Der zweite Eintrag 3-4 gibt z.B. Auskunft darüber, dass die Buchstaben <st> dem phonographischen Prinzip zuzuordnen sind. Zum Schluss erfolgt die Auflistung der Pluralform des jeweiligen Items.

6.3.2 Testaufbau und Tastaturschulung

Im nächsten Schritt wurde das webbasierte ReKoMe konzipiert und entwickelt, dessen Aufbau und Bedienung kindgerecht und motivierend gestaltet, eine leichte Navigation durch den Test erlauben und durch eine Feedbackfunktion motivierend sein sollte. Der Test soll eigenständig von den Schüler*innen durchführbar sein und somit ein praktikables diagnostisches Instrument für den Unterrichtsalltag darstellen. Die Vorerfahrungen der Schüler*innen mit der Arbeit an einem PC sind sehr heterogen. Deshalb wurde eine Tastaturschulung, Tippinstruktion und ein Abtipptest konzipiert, damit ein Umgang mit der Tastatur bzw. der Testumgebung geübt werden kann, wobei ein kleiner animierter Drache Namens Levumi die Kinder über die Sprachausgabe durch das Programm führt. Für einen ersten Prototypen des Programms wurden die Anweisungen von einer Sprachheilpädagogin eingesprochen und später nach einer Erprobung und anschließenden Überarbeitung durch professionelle Tonaufnahmen einer Sprecherin ersetzt.

Es gibt nur bestimmte Möglichkeiten, den Test der Schulpraxis zugänglich zu machen. Der Test kann im Rahmen eines zu installierenden Programms auf einem Datenträger oder digital über eine Lernplattform bereitgestellt werden, wobei die Lernplattform eindeutig Vorteile aufweist (Gebhardt et al., 2016).

Lernplattformen bilden die klassische Hierarchiestruktur der Schule ab und sind eine ideale Möglichkeit, wichtige Funktionsbereiche für Blended-Learning zu implementieren (Maier, 2014). Darüber hinaus können sie niedrigschwellig genutzt werden.

6.3.3 Implementation des ReKoMe auf der Onlinelernplattform Levumi

Das ReKoMe wird auf der Onlinelernplattform „Levumi“ implementiert, kostenlos zur Verfügung gestellt und der Praxis dadurch leicht zugänglich gemacht.

Die Onlinelernplattform „Levumi“[1] ist ein interdisziplinäres Projekt zwischen Informatik und empirischer Bildungs- und Fachdidaktischer Forschung, wurde 2015 von Prof. Dr. Andreas Mühling und Prof. Dr. Markus Gebhardt konzipiert und befindet sich seitdem in fortlaufender Weiterentwicklung (Mühling et al., 2017).

Die Plattform bietet rechnergestützte, kurze und leicht bedienbare Messverfahren an, sogenannte curriculumbasierte Tests, die den Lehrkräften Einblicke in die Lernverläufe

[1] www.levumi.de

ihrer Schüler*innen ermöglichen. Auf Basis der Ergebnisse der Lernverlaufsbeobachtung wird eine passgenaue Förderung möglich. Lernprobleme können frühzeitig identifiziert werden. Darüber hinaus erhält die Lehrkraft ein Feedback zur Wirksamkeit ihres Unterrichts (Gebhardt et al., 2016; Mühling et al., 2017). Entsprechende Handbücher und Videotutorials unterstützen die Lehrkräfte bei der Bedienung und Ergebnisinterpretation der Plattform im Allgemeinen und der Tests im Speziellen und geben Einblicke in die theoretischen Grundlagen. Die Ergebnisse der jeweils verfügbaren Tests der unterschiedlichen Domänen werden sowohl im Kontext aller getesteten Kinder (Klassengraph) als auch auf individueller Ebene (Individualgraph) automatisiert dargestellt. Die Auswertung der Testergebnisse erfolgt sowohl quantitativ als auch qualitativ und ermöglicht auf sehr ökonomische Weise Einblicke in individuelle Lernverläufe. Um die Angebote der Plattform nutzen zu können, sind lediglich ein PC und ein Netzzugang erforderlich. Die Schülerdaten können praktikabel ortsungebunden von einem internetfähigem Gerät jederzeit abgerufen werden (Mühling et al., 2017). Das Anliegen der Plattform ist es zum einen, den Schüler*innen onlinebasierte Lernverlaufstests kostenlos zugänglich zu machen und den Lehrkräften evaluierte Instrumente zur Diagnostik und Förderung an die Hand zu geben. Zum anderen wird eine Plattform angeboten, auf der empirische Daten erhoben werden können, die einen weiteren Beitrag zur Weiterentwicklung geeigneter Diagnoseinstrumenten zur Lernverlaufsbeobachtung leisten (Mühling et al., 2017). Es wird ein Feldzugang für die Lehr-Lernforschung geschaffen, mit dem sehr ökonomisch Daten für die Evaluation neuer Instrumente erhoben werden können und damit der Schulpraxis validierte Instrumente zur Verfügung gestellt werden können (Mühling et al., 2017).

6.4 Konstruierte Instrumente

Im Folgenden werden die aus dem vorherigen beschriebenen Konstruktions- und Operationalisierungsprozess entwickelten Instrumente beschrieben:

- eine papierbasierte Version des ReKoMe,
- das ReKoMe und
- eine Tastaturschulung.

6.4.1 Papierbasierte Version

Für die Pilotierung der operationalisierten Testaufgaben wurde zunächst ein papierbasiertes Rechtschreibkompetenz-Messverfahren (Paper-Pencil Test) entwickelt. Die Testaufgaben bestehen aus 53 Wörtern, die anhand des sprachsystematischen Rechtschreibkompetenzmodells operationalisiert wurden und die den Schüler*innen im Satzkontext im Klassenverbund diktiert werden. Jedes Wort prüft eine oder mehrere Teilkompetenzen des sprachsystematischen Rechtschreibkompetenzmodells. Die Testaufgaben sind identisch mit den Testaufgaben (vgl. Kap. 6.2) des webbasierten ReKoMe.

Für eine objektive Testdurchführung wurden Durchführungshinweise und wörtliche Anweisungen für die Testleiter*innen entwickelt.

6.4.2 Das Rechtschreibkompetenz-Messverfahren (ReKoMe)

Das Rechtschreibkompetenz-Messverfahren (ReKoMe) [2] ist ein webbasiertes Instrument zur differenzierten Lernverlaufsdiagnostik, das ökonomisch im Unterricht zur individuellen Diagnostik eingesetzt werden kann und die sprachlichen Voraussetzungen und heterogenen Lernausgangslagen der Kinder durch die Verwendung eines adäquaten und validierten Rechtschreibkompetenzmodells sowie die Ergebnisse der Sprach- und Schriftlichkeitsforschung berücksichtigt (Blatt et al., 2015; Hinney, 2010). Die Lernentwicklung der Schüler*innen kann mit ReKoMe differenziert und zuverlässig eingeschätzt und Aussagen darüber getroffen werden, wie sich der Rechtschreibunterricht auf die individuelle Kompetenzentwicklung der Kinder auswirkt.
Durch den webbasierten Testmodus ist eine hohe, von der Testleitung unabhängige Durchführungs- und Auswertungsobjektivität gewährleistet und eine ökonomische Testauswertung möglich, da die Antworten direkt erfasst werden können. Die Beantwortung der Testaufgaben erfolgt schriftlich per Tastatur durch die Schüler*innen. Eine nachträgliche manuelle Eingabe der Testergebnisse durch die Lehrkraft ist nicht erforderlich.

Das ReKoMe kann als Einzel- oder Gruppentest eingesetzt werden und ist von den Schüler*innen eigenständig durchführbar. Der Aufbau und die Bedienung des Tests ermöglichen eine einfache und kindgerechte Navigation durch das Programm. Die Instruktionen enthalten Erläuterungen, die klare Handlungsanweisungen zur Beantwortung der Testaufgaben geben. Die Schüler*innen haben außerdem Zugang zu einer Tastaturschulung, die sie durch Erklärungen, Anweisungen und praktische Übungen mit der Tastatur vertraut macht. Dieses Training sollte einmal vor dem ersten Wortdiktat durchgeführt werden und kann bei Bedarf mehrmals wiederholt werden. Die Testantworten werden durch einen eigens für den Test entwickelten Algorithmus automatisiert und differenziert analysiert und codiert. Durch die automatisierte Testauswertung können unmittelbar nach Abschluss des Tests Testberichte und Rückmeldungen bereitgestellt und entsprechende Förderhinweise gegeben werden. Die sofortige Rückmeldung über den Lernstand kann sich unmittelbar positiv auf die Lernentwicklung auswirken. Die Theoretische Grundlage des ReKoMe bildet das sprachsystematische Rechtschreibkompetenzmodell. Es ist geeignet, Unterschiede in den Kompetenzstrukturen von Leistungsgruppen zuverlässig zu identifizieren (Blatt, Prosch & Lorenz, 2016; Jarsinski, 2014; Naujokat, 2015; Voss et al., 2007). Das Modell berücksichtigt die sprachlichen Voraussetzungen der Schüler*innen und setzt anstelle des lautgetreuen Schreibens oder des Regellernens das Erkunden und Verstehen von Schriftstrukturen für den Erwerb von Rechtschreibkompetenz (Blatt et al., 2015; Blatt, Prosch & Frahm, 2016). Das sprachsystematische Rechtschreibkompetenzmodell umfasst nicht

[2] Die Autorin hat den Aufbau und Inhalt des Rechtschreibkompetenz-Messverfahrens bereits unter dem Namen „Rechtschreibtest: Wörter schreiben-Wortdiktat“ in *„Levumi: Handbuch für Lehrkräfte. Deutsch.“* (Jungjohann, Mau, Diehl & Gebhardt, 2019) in ähnlicher Form veröffentlicht. Die Autorin ist die Verfasserin und Verantwortliche aller Texte, die den Themenbereich Rechtschreibung betreffen (vgl. S.4-5; S.15-17).

nur eine Zusammenstellung von orthografischen Regeln auf den Grundlagen der amtlichen Rechtschreibung, sondern auch eine Systematik der Orthografie, die eine Einsicht in die Schriftstruktur der Wörter ermöglicht und die Regularitäten der Wortschreibung erklärbar macht (Eisenberg, 2016). Die sprachsystematische Sichtweise wendet sich vom vorherrschenden Konzept der lautgetreuen Schreibung zum Schriftspracherwerb ab, in dem das Geschriebene als Abbild der gesprochenen Sprache gilt. Der Rechtschreiberwerbsprozess dient nicht nur zur Erfüllung gesellschaftlicher Normen, vielmehr steht das hohe Lernpotenzial dieses Prozesses für die mündliche und schriftliche Sprachkompetenz im Fokus (Eisenberg & Fuhrhop, 2007). Das sprachsystematische Rechtschreibkompetenzmodell benennt fünf Anforderungsbereiche, denen jeweils Teilkompetenzen zugeordnet sind und unterschiedliche Anforderungen an die Schreibenden darstellen.
In den Anforderungsbereichen des Kernbereichs fallen alle Rechtschreibphänomene, die regelbasiert herleitbar und durch einen Wissenstransfer erlernbar sind. Dazu zählen das phonographische, silbische, morphologische, wortübergreifende und Wortbildungs Prinzip. Der Peripheriebereich umfasst die Ausnahmen, die die Kinder sich überwiegend durch Üben einprägen müssen (vgl. Tab. 6.3).

Orientierung an Prinzipien	**Teilkompetenzen**	**Rechtschreibphänomene**
Phonographisches und silbisches Prinzip im Kernbereich	Bezug herstellen zwischen Schrift- und Lautstruktur unter Berücksichtigung der silbenstrukturellen Informationen (Silbenanfangs- und -endrand und Silbenschnitt)	Phonem-Graphem-Korrespondenzen, (Silbenschnitt), Vokallänge, Silbengelenke, Verkürzung des Silbenanfangsrands
Morphologisches Prinzip im Kernbereich	Vererbte silbenschriftliche Informationen in flektierten und abgeleiteten Formen herleiten; Flexionsmorpheme kennen und anwenden	Stammschreibweisen bezogen auf Auslautverhärtungen, Umlautschreibungen und auf Vererbung besonderer Schreibungen (Silbengelenke, silbentrennendes <h>, Vokallänge)
Peripheriebereich	Markierungen in offenen Silben setzten und vererbte Schreibweisen herleiten; Transfer bei Sonderfällen und Lernwörter; Fremdwortschreibung	Doppelvokale, Dehnungs-h
Prinzipien der Wortbildung	Wortarten und Wortbildungsmorpheme kennen und in Ableitungen und Komposita produktiv anwenden	Prä- und Suffixe, Kompositionen Fugenelemente (z.B. s-Fuge)
Wortübergreifendes Prinzip	Syntaxstrukturen kennen und für Groß-, Getrennt- und Zusammenschreibung, dass – Schreibung und Kommasetzung anwenden	

Tabelle 6.3: Erweiterte Tabelle der Rahmenkonzeption zum sprachsystematischen Rechtschreibtest nach Blatt et al. (2011, S.237)

Der Itempool umfasst 53 Testwörter, die speziell auf die Zielgruppe zugeschnitten sind und durch Lupenstellen in der Buchstabenreihenfolge jedes Wortes eine oder mehrere Teilkompetenzen des sprachsystematischen Rechtschreibkompetenzmodells prüfen. Die Lupenstellen, die die jeweiligen Teilkompetenzen repräsentieren, haben in der Gesamtheit des Wortpools eine unterschiedliche Gewichtung (vgl. Tab. 6.2).

Das ReKoMe liegt derzeit für die dritte Klassenstufe vor. Das Verfahren enthält für jeden Anforderungsbereich der Rechtschreibkompetenz eigene Lupenstellen, so dass es auch in niedrigeren und höheren Klassenstufen eingesetzt werden kann. Der inhaltliche Schwerpunkt des ReKoMe liegt auf dem phonographisch-silbischen und morphologischen Prinzip. Um mit dem Verfahren arbeiten zu können, müssen Schüler*innen Kenntnisse zur Graphem-Phonem-Korrespondenz mit silbenstrukturellen Informationen haben sowie in der Lage sein, Ableitungsregeln mit silbenschriftlichen Informationen zur Wortschreibung zu nutzen.

Da der Ausgangswortschatz durch insgesamt fünf Teilkompetenzen des sprachsystematischen Rechtschreibkompetenzmodells eingeteilt und strukturiert ist, kann eine repräsentative Itemstichprobe mittels eines „Zufallsgenerators“ erstellt werden (Wilbert & Linnemann, 2011). Die Grundmengen der Teilkompetenzen sind klar voneinander abgrenzbar und stellen homogene Teilmengen mit jeweils unterschiedlichen Anforderungsbereichen dar. Die Aufgaben werden je Testung und je Kind nach einem proportional-zufälligen Verfahren erzeugt. Dadurch lässt sich eine sehr große Anzahl an unterschiedlichen Testversionen erstellen. Die Gewichtung der einzelnen Teilmengen der Itemstichprobe ist durch die Anzahl der jeweiligen Lupenstellen der Teilkompetenzen im gesamten Itempool bestimmt. Jedes Kind erhält pro Testung eine individuelle Zufallsstichprobe von Testaufgaben, die per Itemsampling erzeugt werden. Die Wörter werden jeweils pro Kind und Messung automatisiert in eine zufällige Reihenfolge gebracht. Abhängig von der individuellen Bearbeitungsgeschwindigkeit der Kinder können unterschiedlich viele Wörter bearbeitet (maximal 53 Wörter) werden. Der Test kann beliebig oft durchgeführt werden. Die Durchführung erfolgt sehr test- und zeitökonomisch durch kurze Wortdiktate am PC. Dabei führt der kleine Drache Levumi die Testpersonen über die Sprachausgabe durch das Programm. Die Wortdiktate werden zuerst im Satzzusammenhang vorgelesen und dann einzeln wiederholt, z.B. für das Wort Blume: „Die Blume ist gelb. Blume“. Die Bearbeitungszeit des Tests beträgt 20 Minuten. Ist die Zeit abgelaufen, beenden die Testpersonen die Eingabe des letzten Wortes.
Erfahrungsgemäß dauert die erste Durchführung einschließlich aller Vorbereitungen für das Wortdiktat und das Tastaturtraining 45 Minuten. Der Test ist für die Tastatur konzipiert, die Maus wird während des Tests nicht benötigt.

Die darauffolgenden Testungen dauern ungefähr 20-30 Minuten (inklusive Vorbereitung), da die Tastaturschulung entfällt und die Schüler*innen mit dem Programm schon vertrauter sind.

Dem Wortdiktat wird ein Beispiel vorangestellt, das von dem Drachen Levumi vorgestellt wird:

> „Hallo, ich bin Levumi. Heute möchte ich gerne wissen, wie gut du schon schreiben kannst. Einige Wörter sind leicht, andere Wörter sind schwer. Bitte konzentriere dich und schreibe so gut, wie du kannst!“ „Bitte drücke jetzt irgendeine Taste auf der Tastatur, damit du beginnen kannst“

Die Schüler werden via Kopfhörer darum gebeten, eine Taste zu drücken, um mit dem Programm zu starten. Die Aufgaben können so individuell angefangen werden. Es folgen zwei Beispiele:

> „Jetzt machen wir zwei Beispiele, damit du weißt, was du aufschreiben sollst. Zuerst lese ich dir einen Satz vor. Danach diktiere ich dir das Wort aus dem Satz, welches du schreiben sollst. Dann schreibst du das Wort auf der Tastatur.“ „Wenn du ein Wort falsch geschrieben hast, kannst du die Leertaste benutzen und die Buchstaben löschen, damit du das Wort noch einmal schreiben kannst. Wenn du das Wort nicht verstanden hast, drücke eine beliebige Taste, dann lese ich dir den Satz und das Wort vor, das du schreiben sollst vor. Achte ganz genau darauf, ob ein Wort groß oder klein geschrieben wird.“ „Das erste Beispiel lautet: Die Blume ist gelb. Blume. Das zweite Beispiel lautet: Der Baum ist groß. Baum. Ich beginne nun, dir die Wörter zu diktieren: Der Hund geht an der Leine. Leine.“

Das Messverfahren verfügt über einen implementierten Algorithmus, der die Ergebnisse automatisch analysiert, auswertet und in Echtzeit in Form von Klassen- und Einzelgraphen darstellt. Die Auswertung umfasst eine quantitative Analyse aller richtig geschriebenen Wortdiktate auf Wortebene. Eine Lösungswahrscheinlichkeit gibt das Verhältnis zwischen richtig und falsch geschriebenen Wörtern bezogen auf die tatsächlich bearbeiteten Wörter an. Die Einteilung in die höchste (viertes Quartil) und die niedrigste Lösungswahrscheinlichkeit (erstes Quartil) gibt Auskunft darüber, welche Wörter über alle Tests hinweg am häufigsten bzw. am seltensten richtig geschrieben wurden. Diese Auswertung steht nach der dritten Durchführung des Verfahrens zur Verfügung.

Im Klassengraphen werden alle Lernverläufe der Testpersonen im Verlauf der einzelnen Testungen im Bezug zur Klassenleistung dargestellt. An den einzelnen Lernverläufen kann die Anzahl der richtig gelösten Testaufgaben der jeweiligen Testpersonen pro Messung abgelesen werden. Im Individualgraphen wird der Lernstand pro Testung der jeweiligen Testperson im Bezug zur Klassenleistung anhand von drei Leistungsquartilen (25, 50, 75) angezeigt. Anhand der Lerngraphen kann auf einem Blick abgelesen werden, ob es Leistungsfortschritte, Leistungsrückschritte oder Stagnationen im Lernverlauf gibt. Zum anderen werden die Ergebnisse auf der Ebene der Teilkompetenzen des sprachsystematischen Kompetenzmodells qualitativ analysiert [3]. Die Darstellung der Auswertung der richtig gelösten Lupenstellen auf Ebene der Teilkompetenzen erfolgt anhand eines Balkendiagramms (vgl. Abb. 6.2).

[3] Die Wortübergreifende Teilkompetenz prüft die Groß- und Kleinschreibung, wurde in der vorliegenden Arbeit bei der Testauswertung auf Grund des webbasierten Testmodus nicht mit einbezogen.

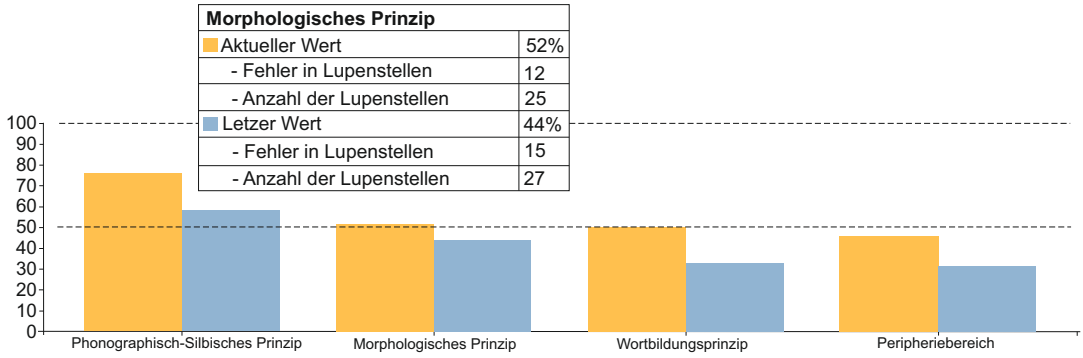

Abbildung 6.2: Auswertung der Lupenstellen auf Ebene der Teilkompetenzen

Der orangefarbene Balken zeigt an, wie viele Lupenstellen im aktuellen Test prozentual richtig beantwortet wurden. Der blaue Balken zeigt, wie viele Lupenstellen beim letzten Test prozentual richtig beantwortet wurden. Zusätzlich gibt die Auswertung Auskunft darüber, wie viele Lupenstellen insgesamt und wie viele davon falsch beantwortet wurden. Die Verteilung der Lupenstellen ist für jeden Test unterschiedlich. Die Lupenstellen, die die jeweiligen Teilkompetenzen repräsentieren, sind im gesamten Wortpool unterschiedlich gewichtet. Der Schwerpunkt des ReKoMe liegt vor allem auf dem phonographisch-silbischen und morphologischen Prinzip.

Die Testpersonen erhalten nach der Eingabe des letzten Testwortes bzw. nach Ablauf der Bearbeitungszeit ein sofortiges Feedback zu ihrer Leistung per Audiodatei anhand der individuellen Bezugsnorm.

6.4.3 Tastaturschulung des ReKoMe

Unter Berücksichtigung der heterogenen Vorerfahrungen im Umgang mit der Tastatur wurde eine Tastaturschulung entwickelt, um den Schüler*innen ein effektives und sicheres Arbeiten während des Tests zu ermöglichen. Die Schulung besteht aus theoretischen Erläuterungen, praktischen Übungen und detaillierten Anleitungen zur Bedienung der Tastatur. Sie sollte einmal vor dem ersten Wortdiktat durchgeführt werden und kann bei Bedarf mehrmals wiederholt werden. Die Durchführungsdauer der Tastaturschulung beträgt ca. 10 Minuten. Aufbau und Inhalt der Tastaturschulung sind im Anhang beschrieben (vgl. Kap. 11.1). Während der Tastaturschulung gibt es immer wieder visuelle Belohnungen in Form eines jubelnden Drachen (Levumi) mit einer Trophäe für die richtige Bearbeitung der Übung. Hat ein*e Schüler*in Probleme bei der Erprobung einzelner Tasten, wird ein Bild mit der Tastenposition eingeblendet und eine Erklärung per Audiodatei eingespielt.

6.5 Zusammenfassung

Das ReKoMe ist ein webbasiertes Testverfahren zur Messung der Rechtschreibkompetenz von Schüler*innen, das auf der Onlinelernplattform Levumi implementiert wurde. Es verfügt über einen entwickelten Algorithmus, der die Testantworten codiert und differenziert analysiert, um die Rechtschreibkompetenzentwicklung einschließlich der jeweiligen Teilkompetenzen zu bestimmen. Der ReKoMe ist in der Lage, differenzierte Lernentwicklungsprofile in Lerngraphen und -balken darzustellen. Aus diesen lassen sich individuelle Förderimplikationen ableiten. Es gilt im Folgenden zu prüfen, inwieweit die Testergebnisse valide Informationen über Leistungsfortschritte, Leistungsrückschritte und Leistungsstagnation enthalten und inwieweit die Ergebnisse des Algorithmus mit denen einer manuellen fachlichen Codierung vergleichbar sind. In zwei kommenden Pilotstudien (vgl. Kap. 8) soll die Qualität des ReKoMe unter Alltagsbedingungen im Schulsystem überprüft werden, um die Integrierbarkeit des RekoMe im Unterricht zu bewerten und mögliche Verbesserungen zu identifizieren. Die anschließende Evaluationsstudie (vgl. Kap. 9) wird sich auf die psychometrische Güte des ReKoMe mittels Analysen der klassischen Testtheorie und der Item-Response-Theorie sowie die Überprüfung der faktoriellen Struktur des zugrunde liegenden theoretischen Rahmenmodells konzentrieren.

7 Methodisches Vorgehen zur Evaluierung des ReKoMe

Das methodische Vorgehen zur Evaluierung des konstruierten ReKoMe wird ausgehend von der Beschreibung des Untersuchungsplans in Kapitel 7.1 dargestellt. Das Kapitel 7.2 beschreibt die Analysemethoden zur Evaluierung des ReKoMe. Im Fokus stehen dabei die Verfahren der Klassischen Testtheorie und der Item-Response-Theorie. Die Anwendung dieser Analyseverfahren ermöglicht die Überprüfung der Gütekriterien des ReKoMe und damit eine fundierte Aussage über die Qualität des Messverfahrens.

7.1 Untersuchungsplan

Im Rahmen einer längsschnittlichen Untersuchung zur Güte des Messverfahrens wird die Rechtschreibkompetenzentwicklung von 146 Schüler*innen acht dritter Klassen in Schleswig-Holstein zu fünf Messzeitpunkten im Zeitraum von September 2017 bis Februar 2018 im vierwöchigen Abstand mit dem ReKoMe erhoben. Für die Durchführung aller Studien in Schleswig-Holstein wurde eine Genehmigung beim Ministerium für Bildung, Wissenschaft und Kultur des Landes Schleswig-Holsteins beantragt und erteilt.

Als Basis für die Konstruktion des ReKoMe erfolgt zunächst die Entwicklung einer papierbasierten Version des Messverfahrens (ReKoMe-PP) auf Basis des sprachsystematischen Rechtschreibkompetenzmodells. Im Rahmen einer Paper-Pencil Studie wird die Rechtschreibkompetenz von 111 Schüler*innen zweier Schulen in Schleswig-Holstein und Bremen zur Pilotierung des Itempools und zur Schaffung einer Datenbasis für die Entwicklung und Überprüfung des Algorithmus mit dem ReKoMe-PP (vgl. Kap. 6.1, 6.3.1) im zweiten Schulhalbjahr 2017 erhoben. Ein webbasierter Prototyp, der auf der Onlinelernplattform Levumi implementiert wurde, wird im zweiten Schulhalbjahr 2017 an drei Schulen in Nordrhein-Westfalen von Schüler*innen der vierten Klassenstufe erprobt (N = 55), um Informationen darüber zu gewinnen, inwiefern sich das Instrument bereits im Unterricht praktikabel integrieren lässt und welche Bereiche einer weiteren Anpassung bedürfen. Für die Überarbeitung des Prototypen werden im Nachgang professionelle Tonaufnahmen erstellt sowie einzelne Anweisungen verkürzt, Instruktionen und Abläufe angepasst.

Die Pilotierung und Evaluation des ReKoMe erfolgt in drei Studien:

> **Paper-Pencil Studie:** Die Paper-Pencil Studie zur Überprüfung des konstruierten Wortmaterials (Itempool) und zur Schaffung einer Datenbasis für die Entwicklung und Überprüfung des Algorithmus fand von März bis April 2017 statt. Die Studie

wurde in sechs dritten Klassen (N = 111) an einer inklusiven Grundschule in Lübeck sowie an einer Grundschule in einem sozialen Brennpunkt in Bremen durchgeführt.

Prototypen Studie: Die empirische Untersuchung zur Erprobung des webbasierten Prototyps des ReKoMe wurde im Juni 2017 in drei vierten Klassen (N = 55) an drei Schulen in Nordrhein-Westfalen (Köln und Dortmund) durchgeführt.

Evaluationsstudie: Die Evaluationsstudie zur Überprüfung der psychometrischen Güte von ReKoMe fand über fünf Messzeitpunkte von Oktober 2017 bis März 2018 in acht dritten Klassen (N = 146) an sechs Schulen in Schleswig-Holstein statt.

Die Stichproben der vorliegenden Untersuchungen sind convenience samples, insgesamt nahmen 17 Grundschulklassen von 12 unterschiedlichen Schulen in Schleswig-Holstein, Bremen und Nordrhein-Westfalen an den Untersuchungen teil. Die teilnehmenden Schulen wurden nicht randomisiert gewählt.

7.2 Datenanalysen

Gemäß der vorgeschlagenen Standardanalysen zur Prüfung der Gütekriterien eines Tests zur Lernverlaufsdiagnostik nach Wilbert und Linnemann (2011) (vgl. Kap. 4.4) werden folgende statistische Analysen auf Basis der Klassischen Testtheorie (vgl. Kap. 4.4.1) und der Item-Response-Theorie (vgl. Kap. 4.4.2) mit dem Programm GNU R vorgenommen:

1. Itemanalysen
2. Dimensionalitätsprüfung
3. Raschanalysen
4. Testfairness

7.2.1 Itemanalysen

Eine erste Beurteilung der Items der jeweiligen Skalen erfolgt anhand von Itemanalysen auf Basis der klassischen Test Theorie (KTT) unter der Verwendung des Pakets *p*sych. Die Ergebnisse zur Itemschwierigkeit, Trennschärfe und internen Konsistenz liefern erste Hinweise, inwiefern sich die Items des Tests zur Abbildung der verschiedenen Merkmalsausprägungen der Testpersonen passen (Kelava & Moosbrugger, 2020a).

Itemschwierigkeit

Items, die im mittleren Schwierigkeitsbereich ($P_i = 50$) liegen, differenzieren am besten zwischen Personen mit höheren und Personen mit niedrigeren Merkmalsausprägungen (Kelava & Moosbrugger, 2020). Um zwischen sehr und weniger leistungsstarken Personen zu differenzieren, sollten die Items im Bereich von $5 \leq P_i \leq 20$ bzw. $80 \leq P_i \leq 95$ liegen. Eine gleichmäßige Verteilung von Items im Schwierigkeitsbereich von $5 \leq P_i \leq 95$ erlaubt eine Differenzierung über das ganze Merkmalsspektrum (Kelava & Moosbrugger, 2020a).

Trennschärfe

Ziel ist es, dass anhand der Items zwischen Personen mit hohen Merkmalsausprägungen und Personen mit niedrigen Merkmalsausprägungen gut differenziert werden kann. Hohe positive Trennschärfen deuten darauf hin, dass die einzelnen Items gut differenzieren. Trennschärfen, die im Bereich von 0.4 bis 0.7 liegen, sind als gut zu bewerten (Kelava & Moosbrugger, 2020a).

Interne Konsistenz

Die interne Konsistenz eines Tests wird im Rahmen der KTT anhand von Kennwerten zu Cronbachs Alpha, zur Retest- und Parallel-Reliabilität beurteilt. Die Schätzung der Reliabilität mit Cronbachs Alpha basiert auf den empirischen Varianzen und Kovarianzen der Itemvariablen. Die Berechnung der Retest- und Parallel-Reliabilität basiert auf den empirischen Korrelationen zwischen den Testwerten paralleler Tests (Schermelleh-Engel & Gäde, 2020).

Verhältnis von Theorie und Empirie

Da die vorliegende Studie einen explorativen Charakter hat und infolgedessen keine a priori Hypothesen formuliert werden können, müssen anhand der Ergebnisse a posteriori Erklärungen aufgestellt werden, aus denen Hypothesen über die Modellverletzungen formuliert werden können (Koller et al., 2012). Eine theoriegeleitete Interpretation der Ergebnisse ist dabei sehr wichtig:

> „Wieso? Ein einfaches Ausscheiden von Items, für die keine theoretische Erklärung gefunden wird, führt zu Theorielosigkeit und artifizieller (künstlicher) Modellanpassung. Es ist einerseits fraglich, ob die Auffälligkeiten bei erneuter Analyse nochmals gefunden werden und andererseits trägt es wenig für zukünftige Itemgenerierungen bei“ (Koller et al., 2012, S.159).

7.2.2 Dimensionalitätsprüfung

Eine konfirmatorische Faktorenanalyse wird durchgeführt, um die Übereinstimmung der Daten mit dem theoretischen Modell zu überprüfen. Mit ihrer Hilfe kann überprüft werden, inwieweit eine Passung zwischen dem theoretischen Modell und den empirischen

Daten besteht und inwieweit die Anzahl der postulierten Faktoren mit der jeweiligen Zuordnung der Items aus dem theoretischen Modell in den Daten repliziert werden kann (Gäde et al., 2020).

Zur Beurteilung der Passung kommen deskriptive- und inferenzstatistische Gütekriterien zur Anwendung, wie z.B. der x^2-Test als inferenzstatistisches Kriterium oder der Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Tucker Lewis Index (TLI) und Standardized Root Mean Square Residual (SRMR) als deskriptive Gütekriterien (Kelava et al., 2020). Es gibt jedoch „kein Gütekriterium, dessen Cut-off-Wert universelle Gültigkeit besitzt, da die Gütekriterien von zahlreichen Einflussfaktoren abhängig sind“ (Gäde et al., 2020, S. 649).

Dazu wird das Paket *l*avaan verwendet. Dabei wird für die Maximum Likelihood-Schätzung ein robuster Standardfehlerschätzer gewählt, da die Daten nicht normalverteilt sind. Um einen Ausschluss von Fällen mit fehlenden Werten bei den Analysen zu vermeiden, wird die «full information maximum likelihood» genutzt.

Zum Vergleich unterschiedlicher Modelle eignen sich das Akaike- (AIC) und Bayesian Information Criterion (BIC) als deskriptives Gütemaß (Gäde et al., 2020). Im Vergleich von zwei Modellen (z.B 1-Faktor vs. 4-Faktoren Modell) zeigt ein kleinerer AIC-und BIC-Wert, dass das Verhältnis von Modellfit und Modellsparsamkeit besser passt als das Vergleichsmodell (Gäde et al., 2020).

7.2.3 Raschanalysen

Das Rasch-Modell wird den Latent-Trait-Modellen zugeordnet und findet eine sehr häufige Verwendung in der Leistungsdiagnostik, deren Ziel es ist, anhand von Testergebnissen auf eine bestimmte Kompetenz (latentes Merkmal) zu schließen (Kelava & Moosbrugger, 2020b). Zur Modellüberprüfung des Rasch Modells -inwiefern die Daten durch das Raschmodell gut beschrieben werden- gibt es unterschiedliche Herangehensweise und „nicht die eine richtige Vorgehensweise“ (Koller et al., 2012, S.157). Vielmehr stellt die Modellüberprüfung „einen kreativen Prozess dar, der oft nicht leicht bewältigbar ist“ (Koller et al., 2012, S.157).

Es liegt ein Datensatz vor, der binär codiert ist. Die Testitems können nur richtig oder falsch gelöst werden. Deshalb wird für die Datenanalysen im Rahmen der IRT ein dichotomes Raschmodell gewählt. Die Schätzung der Itemparameter für das Raschmodell wird mittels des R-Pakets *e*Rm unter der Anwendung der ausschreiben (CML-Methode) in einem ersten Schritt durchgeführt und anhand der Infit-Werte beurteilt. Diese sind im Gegensatz zu den Outfit-Werten im mittleren Leistungsniveau zuverlässiger und weniger anfällig für Ausreißer in den Werten (Linacre et al., 2002). Bei der Itemselektion wird auch das zugrunde liegende theoretische Modell des sprachsystematischen Kompetenzmodells (Blatt et al., 2015) in den Blick genommen, um dem Anspruch eines iterativen Prozesses einer gelingenden Auseinandersetzung von Theorie und Empirie zu genügen (Kelava & Moosbrugger, 2020a). Infit-Statistiken geben darüber Auskunft, inwiefern die jeweiligen Items personenübergreifend zum Messmodell passen. Viele Raschanalyseprogramme

berechnen hierzu den gewichteten Mean-Square der Infit- und Outfit Werte, dessen Erwartungswert 1 ist (Bond et al., 2020). Werte des gewichteten Mean-Squares im Bereich von $.75 \leq \text{Infit} \leq 1.3$ liegen im akzeptablen Bereich (Bond et al., 2020). Ein Infit-Wert von 1.3 gibt z.B. 30% mehr Variation in den beobachteten Daten an als es das Rasch-Modell vorhergesagt hat und ein Infit-Wert von .78 sagt dementsprechend 22% weniger Variation vorher (Bond et al., 2020, S.241). Liegt der Wert der Infit-Statistiken über 1, spricht man von einer durchschnittlich schlechten und einem kleineren Wert unter 1 von einer durchschnittlich besseren Passung. Die Passung der Items hat Einfluss auf die Trennschärfe, die ein Indiz dafür ist, inwiefern ein Item zwischen verschiedenen Eigenschaftsausprägungen der Personen trennt (Rost, 2004).

Danach erfolgt die Schätzung der Personenparameter mit der ausschreiben (ML-Methode). Während im Rahmen der KTT die Reliabilität der messfehlerbehafteten Testvariablen bestimmt wird, wird die Reliabilität der geschätzten latenten Personenwerte im Kontext der IRT modellbasiert bestimmt (Schermelleh-Engel & Gäde, 2020). Daraus kann die Information abgelesen werden, inwieweit die Unterschiede zwischen den geschätzten Personenwerten mit tatsächlichen Unterschieden zwischen den Personen zusammenhängen (Schermelleh-Engel & Gäde, 2020). Die Interpretation der Werte erfolgt analog zur klassischen Testtheorie. Die Beurteilung der übergeordneten Messgenauigkeit der Skalen wird mit dem Paket Tam berechnet und als EAP und WLE Reliabilität berichtet. Personenparameter und Itemparameter können mittels Raschanalysen auf der gleichen Skala abgebildet werden. Dies erlaubt eine Beurteilung der Testaufgaben bezüglich der Angemessenheit der Schwierigkeitsgrade. Die Verteilung der Personenparameter und die Position der Items auf der latenten Dimension werden anhand einer Person-Item-Map dargestellt. Eine Verteilung der Itemschwierigkeiten über den ganzen Bereich der Personenfähigkeit (Latent Dimension) stellt den besten Fall dar (Koller et al., 2012).

Der bedingte Likelihood-Quotienten-Test nach Andersen (1973) eignet sich zur Überprüfung der Itemhomogenität. Dazu wird geprüft, ob sich die Itemparameter aufgeteilt in zwei Gruppen unterscheiden. Treten zwischen den beiden Gruppen keine systematischen Unterschiede auf, kann die Nullhypothese aufrechterhalten werden und von einer Modellkonformität ausgegangen werden (Kelava & Moosbrugger, 2020a). Als Trennkriterium wird der Mittelwert, Median, gewählt. Für die parametrische Modellüberprüfung werden Alpha-Korrekturen vorgenommen.

Modellvergleich

Inwieweit ein 2PL-Modell mit frei schätzbaren Ladungen (Birnbaum-Modell) die Daten signifikant besser erklärt als ein Modell mit identischen Ladungen (Rasch-Modell), wird mit einem Likelihood-Ratio-Test geprüft. Dazu werden die Modellparameter berechnet und anschließend mit einer ANOVA ermittelt, welches Modell die Daten besser erklärt. Für die Schätzung der Parameter des zweiparametrischen logistischen Modells wird das Paket *l*tm verwendet. Als deskriptives Gütekriterium wird das *B*ayesian Information Criterion (BIC) verwendet, da es die Anzahl der Parameter und den Stichprobenumfang berücksichtigt (Rost, 2004). Im Vergleich von zwei Modellen zeigt ein kleinerer BIC-Wert, dass

das Verhältnis von Modellfit und Modellsparsamkeit besser passt als das Vergleichsmodell (Gäde et al., 2020).

7.2.4 Testfairness

Testfairness liegt vor, „wenn die resultierenden Testwerte zu keiner systematischen Benachteiligung bestimmter Personen aufgrund ihrer Zugehörigkeit zu ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppen führen“(Moosbrugger & Kelava, 2020, S.25). Personen mit gleichem Fähigkeitsniveau, die verschiedenen Gruppen angehören, erzielen also in einem Test die gleichen Personenwerte (Koller et al., 2012). Die Prüfung der Subgruppeninvarianz anhand externer Teilungskriterien (Geschlecht) wird als DIF-Prüfung bezeichnet (Koller et al., 2012). Ist ein Item für Personen mit gleicher Fähigkeit unterschiedlich schwer zu lösen, liegt Differential Item Functioning (DIF) vor. Dies kann ein Hinweis darauf sein, dass mit dem Item eine andere Eigenschaft, z.B. die Sprachkompetenz, gemessen wird und die Validität des Items nicht gegeben ist (Wilbert & Linnemann, 2011). Die Methode des DIF lässt auch Aussagen zur Konstruktvalidität eines Tests zu (Wilbert & Linnemann, 2011).

8 Pilotierung des ReKoMe

In diesem Kapitel werden die Ergebnisse der durchgeführten Pilotstudien vor dem Hintergrund der Beschreibung der Stichprobenkonstruktion, der Untersuchungsdurchführung und der angewandten statistischen Analysen vorgestellt. Ziel der Pilotstudien ist, es die Qualität der konstruierten Testaufgaben, des Algorithmus und des webbasierten Testdesigns unter Alltagsbedingungen im System Schule zu überprüfen. Die Ergebnisse der Studien sollen auch dazu beitragen, die Integrierbarkeit von ReKoMe in den Unterricht zu überprüfen und Hinweise auf mögliche Verbesserungen zu identifizieren. Die Ergebnisse der Paper-Pencil Studie werden in Kapitel 8.1 präsentiert und dienen als Datengrundlage für die Entwicklung und Prüfung des Algorithmus des ReKoMe (vgl. Kap. 8.2). Im Anschluss stehen die Ergebnisse der Prototypenstudie im Fokus (vgl. Kap. 8.3). Kapitel 8.5 vergleicht abschließend die Ergebnisse der Paper-Pencil Studie mit den Ergebnissen der Prototypenstudie und zeigt die Überarbeitungsnotwendigkeiten auf.

8.1 Paper-Pencil Studie

Die Pilotierung der konstruierten Testaufgaben erfolgt im Rahmen einer Paper-Pencil Studie in sechs dritten Klassen an zwei inklusiven Grundschulen in Lübeck und Bremen im Zeitraum von März bis April 2017.

8.1.1 ReKoMe - PP

Das Rechtschreibkompetenz-Messverfahren - Paper-Pencil (ReKoMe-PP) beinhaltet 53 Wörter, die den Schüler*innen im Klassenverband im Satzkontext diktiert werden. Die Testaufgaben wurden nach dem sprachsystematischen Rechtschreibkompetenzmodell gemäß dem in Kapitel 6.2 beschriebenen Verfahren operationalisiert und sind mit den Testaufgaben des ReKoMe identisch. Für eine objektive Testdurchführung werden Durchführungshinweise und standardisierte Anweisungen erstellt.

8.1.2 Stichprobenkonstruktion

Basis für die empirische Untersuchung zur Pilotierung der Testaufgaben bildet eine Stichprobe von Schüler*innen (N=111, 55% weiblich) aus sechs dritten Klassen aus zwei Schulen in Lübeck und Bremen im durchschnittlichen Alter von 9.3 Jahren. Die Stichprobe ist ein convenience sample.

8.1.3 Untersuchungsdurchführung

Die Rechtschreibkompetenz der Schüler*innen wird mit dem ReKoMe-PP zu einem Messzeitpunkt im Zeitraum von März bis April 2017 erhoben. Die Erhebung erfolgt sowohl durch die Autorin als auch durch Lehrkräfte. Die Teilnahme an der Studie war freiwillig, es wurde eine Einverständniserklärung der Erziehungsberechtigten aller teilnehmenden Schüler*innen eingeholt.

8.1.4 Datenanalysen

Im Folgenden werden Analysen auf Basis der Klassischen Testtheorie auf Ebene des ganzen Wortes und auf Ebene der Teilkompetenzen durchgeführt. Die Skala...

- Quan misst die Fähigkeit, Wörter auf Ganzwortebene orthografisch richtig schreiben zu können.
- PhonSilb misst die Fähigkeit, einen Bezug zwischen Schriftstruktur und Lautstruktur unter Berücksichtigung der silbenstrukturellen Informationen (Silbeanfangsrand, Silbenendrand und Silbenschnitt) herstellen zu können.
- Morph erfasst die Fähigkeit, vererbte silbenschriftliche Informationen in flektierten und abgeleiteten Formen herleiten zu können und die richtige Anwendung von Flexionsmorphemen.
- Peri erfasst die Fähigkeit, Markierungen in offenen Silben setzen und vererbte Schreibweisen herleiten zu können und Lernwörter und Fremdwortschreibungen richtig zu schreiben.
- Wortbil erfasst die Fähigkeit, Wortarten und Wortbildungsmorpheme zu kennen und in Ableitungen und Komposita richtig anzuwenden.

Anhand der Werte zur Itemschwierigkeit, Trennschärfe, interne Konsistenz und Retestreliabilität lassen sich erste Schlussfolgerungen ziehen, inwiefern die Items des Tests zur Abbildung der verschiedenen Merkmalsausprägungen der Testpersonen passen (Kelava & Moosbrugger, 2020a).

8.1.5 Ergebnisse

Tabelle 8.1 stellt die Ergebnisse der Itemanalysen der Skalen Quan, Phon-Silb, Morph, Peri und Wortbil auf Basis der Klassischen Testtheorie dar.

	Quan	PhonSilb	Morph	Peri	Wortbil
Itemschwierigkeit					
M (SD)	.65 (.24)	.84 (.19)	.71 (18)	.58 (27)	.69 (15)
min-max	.05 - .98	.17 - .99	.13 - .98	.17 - .92	.50 - .90
Trennschärfe					
M(SD)	.42 (.16)	.27 (.14)	.39 (.17)	.41 (.11)	.25 (.18)
min-max	-.02 - .67	-.03 - .54	.05 - .68	.20 - .64	-.14 - .45
interne Konsistenz					
α	.93	.81	.87	.80	.50
Split-Half	.94	.88	.88	.75	.83

Tabelle 8.1: Itemanalysen KTT - Paper-Pencil Studie

Die Items der Skala Quan liegen im mittleren bis leichten Bereich. Die Wörter des Wortdiktats wurden von 65% der Schüler*innen orthografisch korrekt verschriftet.

Das Anforderungsniveau der Lupenstellen des phonographisch-silbischen Prinzips (M = .84, SD = .19), morphologischen (M = .71, SD = .18) und Wortbildungsprinzip (M = .69, SD = .15) lag eher im sehr leichten und die des Peripheriebereichs (M = .58, SD = .27) im mittleren Bereich. Die Anzahl der korrekt verschriftlichten Items auf Ebene der Subskalen von Phon-Silb mit 84% zu Morph 71% über Wortbil 69% und Peri 58% ist theoriekonform. Die Items der Skalen Quan, Morph und Peri differenzieren im Durchschnitt gut zwischen leistungsschwachen und leistungsstarken Schüler*innen. Die Lupenstellen des phonographisch-silbischen Prinzips wurden insgesamt von 84% der Kinder richtig beantwortet und schienen für diese Stichprobe zu einfach gewesen zu sein. Dies lässt auch die geringe Trennschärfe im Mittel von M_rit = .27 erklären. Die Items des Wortbildungsprinzips trennen M_rit = .25 nicht ausreichend zwischen den unterschiedlichen Anforderungsniveaus. Dieses Ergebnis muss vor dem Hintergrund der Anzahl der Lupenstellen (N = 8) jedoch vorsichtig interpretiert werden, da sich im Vergleich zu den anderen Skalen weniger Items im Itempool befinden. Die Anzahl der Testitems hat Einfluss auf die Höhe des Cronbachs Alpha (Schermelleh-Engel & Gäde, 2020).

Auf Ganzwortebene misst die Skala Quan mit einem Cronbachs Alpha von .93 exzellent. Die Skalen Phons-Silb, Morph und Peri messen mit einem Cronbachs Alpha von .80-.87 in einem guten Bereich. Die Skala Wortbil misst nicht zufriedenstellend mit einem Cronbachs Alpha von .50. Die Split-Half Reliabilitäten aller Skalen liegen mit Ausnahme des Peripheriebereichs (.75) zwischen .94 und .82 im exzellenten bis guten Bereich.

8.2 Prüfung des Algorithmus

Die Auswertung von Schriftlösungen auf der Basis des sprachsystematischen Rechtschreibkompetenzmodells ist komplex, zeitaufwendig und erfordert ein hohes Fachwissen. Der Algorithmus basiert auf einem Kategoriensystem zur differenziellen Analyse der Schriftlösungen, anhand dessen auf Basis des sprachsystematischen Rechtschreibkompetenzmodells Struktureinheiten und Ausschlüsse der Wörter definiert werden. Es ist wichtig zu prüfen, ob der im ReKoMe implementierte Algorithmus zuverlässig die Testergebnisse des ReKoMe automatisiert codiert und analysiert.

8.2.1 Datengrundlage

Für die Überprüfung des Algorithmus und des Itempools wird eine Datengrundlage von unterschiedlichen Fehlschreibungen von Schüler*innen benötigt. Basis für die Überprüfung des Algorithmus bilden die im Rahmen der Paper-Pencil Studie zur Pilotierung der Testaufgaben erhobenen Daten. Im Anschluss der Testungen werden die Testergebnisse der Schüler*innen mit den jeweiligen Schreiblösungen in Excel übertragen und nach den grundlegenden Kenntnissen der Sprachwissenschaften anhand des Sprachsystematischen Kompetenzmodells durch die Autorin qualitativ ausgewertet und manuell codiert. Durch den Vergleich der manuellen Codierung der Testergebnisse mit der automatisierten Codierung lässt sich der entwickelte Algorithmus überprüfen. Es liegen insgesamt 833 unterschiedliche Schreiblösungen von insgesamt 54 Testitems [1] vor. Für die unterschiedlichen Wörter liegen jeweils 2-69 verschiedene Schreibweisen vor wie z.B. für das Wort <Teller>: Tehler, Teler, Tella usw., welche die Überprüfung der Korrektheit des Algorithmus ermöglichen.

8.2.2 Ergebnisse

Die Ergebnisse zeigen, dass mit dem Algorithmus die Testergebnisse des ReKoMe zu 99% richtig analysiert und kategorisiert werden. Der Algorithmus prüft zuverlässig und automatisiert zum einem auf Wortebene, wie viele Wörter insgesamt richtig und wie viele Wörter falsch geschrieben wurden. Zum anderen lässt sich mit dem Algorithmus auf Ebene der Teilkompetenzen des sprachsystematischen Kompetenzmodells die Art der Schreiblösungen innerhalb eines Wortes näher analysieren, um differenzierte Lernentwicklungsprofile erstellen zu können, anhand derer sich gezielte Förderimplikationen ableiten lassen.

8.3 Prototypenstudie

Die Pilotierung des webbasierten Testdesigns erfolgt im Rahmen einer empirischen Studie in drei vierten Grundschulklassen an drei Schulen in Köln und Dortmund im Juni 2017.

[1] Ein Testitem stellt ein Beispielwort dar und wurde in den späteren Testpool nicht mit aufgenommen.

Ziel der Studie ist es auch, die Verständlichkeit der Testaufgaben zu überprüfen und inhaltliche sowie technische Schwierigkeiten zu identifizieren.

8.3.1 ReKoMe - Prototyp

Für das webbasierte ReKoMe wird zunächst ein Prototyp entwickelt. Der Prototyp besteht aus einer Tastaturschulung und aus Wortdiktaten (vgl. Kap. 6.2). Der kleine Drache Levumi führt die Schüler*innen über die Sprachausgabe durch das Programm und den Test.

8.3.2 Stichprobenkonstruktion

Basis für die empirische Untersuchung zur Pilotierung des ReKoMe bildet eine Stichprobe von Schüler*innen (N = 55; weiblich = 50,9%) im durchschnittlichen Alter von 10.2 Jahren. Die Stichprobe ist ein convenience sample.

8.3.3 Untersuchungsdurchführung

Die Rechtschreibkompetenz der Schüler*innen wird im Rahmen einer Masterarbeit an der Technischen Universität Dortmund mit dem webbasierten Prototypen des ReKoMe zu einem Messzeitpunkt im Juni 2017 erhoben.

Die Erhebung erfolgte durch eine Studentin. Zur Wahrung der Objektivität der Testung wurden standardisierte Anweisungen zur Testdurchführung konzipiert.

8.3.4 Datenanalysen

Die Testantworten der Schüler*innen werden automatisiert auf Ganzwortebene und auf Ebene der Teilkompetenzen im Prototypen analysiert, codiert und auf der Onlineplattform Levumi gespeichert. Anschließend können die Daten für statistische Analysen exportiert werden. Durch eine Aktualisierung der Onlinelernplattform im Zuge der neuen Datenschutzgrundverordnung im Jahr 2017 ging der Datensatz der Pilotstudie verloren, bevor alle Ergebnisse exportiert wurden. Die deskriptiven Analysen auf Basis der Klassischen Testtheorie können deshalb lediglich auf Ebene des ganzen Wortes durchgeführt werden. Im Folgenden werden die Ergebnisse der Analysen zur Itemschwierigkeit, Trennschärfe und internen Konsistenz dargestellt.

8.3.5 Ergebnisse

Die Schüler*innen dieser Stichprobe verschriftlichten bei der webbasierten Testung 50% der Testitems richtig. Die Items sind durchschnittlich ($M = .50$, $SD = .20$) mittelschwer und trennen sehr gut ($M_rit(SD) = .79$ (.12)) zwischen leistungsschwachen und leistungsstarken Schüler*innen. Die Skala Quan misst die Rechtschreibkompetenz auf Ganzwortebene mit einem Cronbachs Alpha von .99 und einer Split-Half Reliabilität von 1 exzellent.

8.4 Verständlichkeitsanalyse

Um inhaltliche, praktische und technische Schwierigkeiten des ReKoMe zu identifizieren, wurde im Rahmen der Pilotstudien auch Feedback von Lehrer*innen und Schüler*innen zu den Testaufgaben und zum Testdesign eingeholt. Die Lehrer*innen bewerteten das Testdesign als intuitiv und leicht durchführbar. Die Schüler*innen gaben eine positive Rückmeldung zum ReKoMe. Die Anweisungen des kleinen Drachen Levumi wurden von den Schüler*innen als besonders motivierend empfunden. Die Durchführungshinweise während der Tastaturschulung und des Wortdiktats erwiesen sich teilweise als zu lang und komplex und wurden gekürzt wie z.B. die Anweisungen zur Bedienung der Löschtaste:

Alt: „Bestimmt hast du schon einmal an einem Computer etwas geschrieben, das du wieder löschen wolltest. Dafür benutzt man die Löschtaste. Auf dem Bild kannst du erkennen, wo du die Taste findest. Bitte drücke jetzt die Löschtaste auf der Tastatur.“

Neu: „Wenn du Buchstaben löschen möchtest, benutze die Löschtaste. Auf dem Bild kannst du erkennen, wo du die Taste findest. Bitte drücke jetzt die Löschtaste.“

Die Audioaufnahmen einzelner Wörter waren hinsichtlich der Aussprache und Betonung manchmal nicht präzise, wie zum Beispiel beim Wort „bissig“. Die Audiodatei für dieses Testwort hat zu vielen Fehlschreibungen geführt und wurde deshalb von einer professionellen Sprecherin erneut aufgenommen. Weitere mögliche Problemfelder, die sich aus dem Erhebungsmodus mit einem PC ergeben, wurden identifiziert und in die Durchführungshinweise mitaufgenommen. Zur Überarbeitung des bereits skizzierten Prototyps wurden im Nachhinein professionelle Tonaufnahmen angefertigt und einzelne Anweisungen gekürzt sowie Anweisungen und Abläufe angepasst.

8.5 Zusammenfassung

Die Ergebnisse der Pilotstudien zeigen, dass sich die operationalisierten Testitems der Skala Quan sowohl für die papierbasierte als auch für die webbasierte Messung exzellent eignen, um Rechtschreibkompetenz auf Ganzwortebene valide zu messen. Die im Rahmen der Paper-Pencil Studie pilotierten Testitems auf der Ebene der Teilkompetenzen

des phonographisch-silbischen Prinzips, morphologischen Prinzips und des Peripheriebereichs haben eine gute interne Konsistenz und eignen sich gut zur Erfassung der differenzierten Rechtschreibkompetenzentwicklung. Für die Skala Wortbildungsprinzip ist die interne Konsistenz nicht zufriedenstellend. Hervorzuheben ist, dass der entwickelte Algorithmus mit einer Wahrscheinlichkeit von 99% die Schreiblösungen richtig quantitativ und qualitativ codiert, analysiert und auswertet. Damit ist die Grundlage für eine qualitative Ergebnisanalyse geschaffen, aus der zuverlässig, schnell, effektiv und ressourcenorientiert Förderimplikationen abgeleitet werden können. Die Ergebnisse der Paper-Pencil Testung und der webbasierten Testung lassen sich nicht zuverlässig miteinander vergleichen, da die Stichproben unterschiedlich groß sind und aus verschiedenen Bundesländern und Klassenstufen stammen. Im Durchschnitt wurden in der Paper-Pencil Testung 65% und in der webbasierten Testung 50% aller Wörter auf Ganzwortebene richtig geschrieben. Obwohl die webbasierte Testung in der vierten Klassenstufe durchgeführt wurde, erzielten die Schüler*innen durchschnittlich 15% weniger richtig verschriftete Wörter als bei der Paper-Pencil Testung in der dritten Klassenstufe. Eine vorsichtige Interpretation könnte sein, dass bei der webbasierten Testung mehrere Kompetenzen, wie das Tippen auf der Tastatur, gleichzeitig gefordert werden und dass die unterschiedliche Anzahl von Schüler*innen mit Migrationshintergrund oder sonderpädagogischem Förderbedarf in der Stichprobe ebenfalls einen Einfluss auf die Testergebnisse haben könnte.

Nachdem die Testkonstruktion und der vorläufige Itempool in zwei Pilotstudien auf ihre Güte hin überprüft und im Rahmen der qualitativen Verständlichkeitsprüfung angepasst wurden, werden im Folgenden die Ergebnisse der Evaluationsstudie vorgestellt, um die Passung der Items mit dem zugrunde gelegten psychometrischen Modell mit Analysen der Item-Response-Theorie zu überprüfen.

9 Evaluation des ReKoMe

Im Mittelpunkt dieses Kapitels stehen die Ergebnisse der Evaluationsstudie des ReKoMe (vgl. Kap. 9.3), die Beantwortung der Forschungsfragen der Arbeit (vgl. Kap. 9.9) und die Darstellung von Fallbeispielen (vgl. Kap. 9.10). Die Studie konzentriert sich auf die Überprüfung der psychometrischen Güte des konstruierten Rechtschreibkompetenz-Messverfahrens mittels Analysen der klassischen Testtheorie und der Item-Response-Theorie sowie die Überprüfung der faktoriellen Struktur des zugrunde gelegten theoretischen Rahmenmodells. Die theoretisch postulierte faktorielle Struktur des sprachsystematischen Rechtschreibkompetenzmodells wurde in der dritten Klassenstufe bisher noch nicht empirisch überprüft. Die Ergebnisse der Evaluationsstudie werden im Kontext der Beschreibung der Stichprobenkonstruktion (vgl. Kap. 9.1), der Untersuchungsdurchführung (vgl. Kap. 9.2) und der angewandten statistischen Analysen (vgl. Kap. 9.3) vorgestellt.

9.1 Stichprobenkonstruktion

Die Evaluationsstudie wird im längsschnittlichen Design unter Alltagsbedingungen im System Schule durchgeführt. Die Erhebung der Rechtschreibkompetenzentwicklung mit dem ReKoMe von insgesamt 146 Drittklässler*innen in Schleswig-Holstein erfolgt im zweiten Schulhalbjahr in einem vierwöchigen Abstand von Oktober 2017 bis März 2018 zu fünf Messzeitpunkten in insgesamt acht Klassen. Für die Durchführung der Studie wurde 2016 beim Ministerium für Bildung, Wissenschaft und Kultur des Landes Schleswig-Holsteins ein Antrag gestellt. Die Genehmigung wurde 2017 erteilt. Computerarbeitsplätze mit Internetzugang sind eine Grundvoraussetzung für den Einsatz von ReKoMe. Für die Rekrutierung der Stichprobe stellt dies eine besondere Herausforderung dar, da die Ausgangssituation der einzelnen Schulen diesbezüglich sehr heterogen ist. Die technische Ausstattung und die Sicherstellung einer zuverlässigen Internetverbindung sowie das Vorhandensein einer geeigneten räumlichen Ausstattung sind nicht an allen Grundschulen gegeben. Zur Stichprobenrekrutierung wurde eine Internetrecherche durchgeführt mit dem Ziel, Grundschulen in Schleswig-Holstein zu identifizieren, die über die notwendige technische Infrastruktur zur Durchführung der Studie verfügen. Dazu wurden die Homepages der Schulen in Schleswig-Holstein nach entsprechenden Informationen analysiert. Im Anschluss an die Analyse wurden die Schulleitungen von insgesamt 20 Grundschulen in Schleswig-Holstein in einem Anschreiben über das Forschungsvorhaben informiert und um ihr Einverständnis zur Durchführung der Studie an ihrer Schule gebeten. Da der Rücklauf sehr gering war, von insgesamt 20 angeschriebenen Schulleitungen meldete sich nur eine

Schule, wurden die Schulleitungen zusätzlich telefonisch von der Autorin kontaktiert und über die geplante Studie informiert. Alle Schulleiter*innen, zuständigen Lehrkräfte und Erziehungsberechtigten wurden mit umfangreichem Informationsmaterial zum Inhalt und zur Durchführung der Studie versorgt. Den Schulleitungen wurde zugesichert, dass bei Teilnahme an der Studie auch eine umfassende detaillierte Ergebniszusammenfassung zu den Rechtschreibleistungen der Schüler*innen erfolgt und den jeweiligen Deutschlehrkräften zur Verfügung gestellt wird. Von allen Erziehungsberechtigten wurde das Einverständnis zur Teilnahme an der Studie eingeholt. Zunächst konnten insgesamt acht Klassen aus drei Grundschulen für die Teilnahme an der Studie gewonnen werden. Aus technischen Gründen fiel jedoch kurzfristig eine Schule mit insgesamt vier Grundschulklassen aus, da sich bei der ersten Erhebung zeigte, dass die Internetverbindung in dieser Schule nicht stabil war. Somit halbierte sich die Stichprobe kurz vor Beginn der Studie. Zur Vergrößerung der Stichprobe wurden daher vier Lehramtsstudierende im Rahmen von Bachelorarbeiten an der Datenerhebung beteiligt und deren Kontakte zu weiteren Grundschulen im Raum Schleswig-Holstein genutzt. Die Studierenden wurden von der Autorin umfassend im Thema Lernverlaufsdiagnostik im Bereich Rechtschreibung und für die Durchführung der Studie als Testleiter*innen ausgebildet. Die Rekrutierung von vier weiteren Schulen erfolgte durch die Studierenden, wobei Informationsmaterialien sowie Schul- und Elternanschreiben von der Autorin zur Verfügung gestellt wurden. Insgesamt konnten acht dritte Grundschulklassen aus sechs verschiedenen Schulen in Schleswig-Holstein für die Teilnahme an der Evaluationsstudie gewonnen werden. Zwei Schulen liegen südöstlich von Kiel, eine im Kreis Schleswig-Flensburg, eine im Kreis Dithmarschen und eine im Süden Schleswig-Holsteins im Kreis Stormarn. Die Größe der Schulen ist unterschiedlich. Die kleinste Schule besteht aus 97 Schüler*innen und die größte aus 700 Schüler*innen. Zwei Schulen haben einen hohen Anteil von Schüler*innen mit Migrationshintergrund. Davon hat eine Schule ein angeschlossenes DaZ-Zentrum. Eine weitere Schule hat ein angeschlossenes Förderzentrum.

9.2 Untersuchungsdurchführung

Das ReKoMe wird zu fünf Messzeitpunkten im Zeitraum von September 2017 bis Februar 2018 im vierwöchigen Abstand zur längsschnittlichen Erfassung der Rechtschreibkompetenzentwicklung in den jeweiligen Klassen der Stichprobe parallel eingesetzt. Die Datenerhebung erfolgt durch die Autorin und durch Bachelor-Studentinnen (vgl. Kap. 9.1). Zur Durchführung des ReKoMe erfolgen auf eine sehr ökonomische Weise kurze Wortdiktate am PC. Dabei führt der kleine Drache Levumi die Testpersonen über die Sprachausgabe durch das Programm. Die Testwörter werden zuerst im Satzzusammenhang vorgelesen und dann einzeln wiederholt, z.B. für das Wort Blume: „Die Blume ist gelb. Blume“. Der Test dauert 20 Minuten. Ist die Zeit abgelaufen, beenden die Schüler*innen die Eingabe des letzten Wortes. Erfahrungsgemäß dauert die erste Durchführung einschließlich aller Vorbereitungen für das Wortdiktat und das Tastaturtraining 45 Minuten. Der Test ist für die Tastatureingabe konzipiert, die Maus wird während des Tests nicht betätigt oder berührt. Erfahrungsgemäß dauert die erste Durchführung inklusive aller Vorbereitungen

des Wortdiktats und der Tastaturschulung 45 Minuten. Der Test ist für die Tastatur konzipiert, die Maus wird während der Testung nicht betätigt oder berührt. Die darauffolgenden Testungen dauern ungefähr 20-30 Minuten inklusive der Vorbereitungen, da die Tastaturschulung entfällt und die Schüler*innen mit dem Programm schon vertrauter sind.

Da der Ausgangswortschatz durch insgesamt fünf Teilkompetenzen des sprachsystematischen Rechtschreibkompetenzmodells eingeteilt und strukturiert ist, kann eine repräsentative Itemstichprobe mittels eines „Zufallsgenerators“ erstellt werden (Wilbert & Linnemann, 2011). Die Grundmengen der Teilkompetenzen sind klar voneinander abgrenzbar und stellen homogene Teilmengen mit jeweils unterschiedlichen Anforderungsbereichen dar. Die Aufgaben werden je Testung und je Kind nach einem proportional-zufälligen Verfahren erzeugt. Dadurch lassen sich eine extrem große Anzahl an unterschiedlichen Testversionen erzeugen. Die Gewichtung der einzelnen Teilmengen der Itemstichprobe ist durch die Anzahl der jeweiligen Lupenstellen (vgl. Kap. 6.4.2) der Teilkompetenzen im gesamten Wortpool bestimmt. Jedes Kind erhält pro Testung eine individuelle Zufallsstichprobe von Testaufgaben, die per Itemsampling erzeugt werden. Die Wörter werden jeweils pro Kind und Messung automatisiert in eine zufällige Reihenfolge gebracht. Abhängig von der individuellen Bearbeitungsgeschwindigkeit der Kinder können unterschiedlich viele Wörter bearbeitet (maximal 53 Wörter) werden (vgl. Kap. 6.4.2). Der Test kann beliebig oft durchgeführt werden. Die Tastaturschulung soll vor dem ersten Wortdiktat mindestens einmal erfolgt sein und kann bei Bedarf mehrmals wiederholt werden. Die Durchführungsdauer der Tastaturschulung beträgt ca. 10 Minuten. Da zum ersten Erhebungszeitpunkt vor dem Wortdiktat des ReKoMe eine Tastaturschulung erfolgt, deren Durchführung ca. 10 Minuten beträgt und die Schüler*innen noch keine Erfahrungen mit der Handhabung des Programms haben, wird die Bearbeitungszeit des ReKoMe einmalig auf 15 Minuten begrenzt. Ab dem zweiten Erhebungszeitpunkt beträgt die Bearbeitungszeit 20 Minuten. Die Durchführung der Testung inklusive aller Vorbereitungen dauert ca. 45-60 Minuten pro Klasse. Die Testdurchführung muss an die technischen und organisatorischen Rahmenbedingungen der Schulen angepasst werden. Zum Zeitpunkt der Erhebung verfügen die Schulen über keine ausreichenden Computerarbeitsplätze, die eine gleichzeitige Testung aller Schüler*innen möglich macht. Dadurch entsteht ein erheblicher organisatorischer Aufwand, der teilweise auch Unruhe bei den Schüler*innen hervorruft. So befanden sich z.B. an einer Schule alle Schüler*innen in einem Raum und wechselten sich nacheinander ab. Teilweise störten sich die Schüler*innen dabei gegenseitig. An anderen Schulen verlief dieser Wechsel aber auch sehr diszipliniert. Zudem zeichnete sich ein sehr differenziertes Bild im sicheren Umgang mit einem Computer ab. An einigen Schulen war die Arbeit am Computer trotz Computerarbeitsplätzen kaum Gegenstand des Unterrichts. Dadurch wurde sehr viel Hilfestellung im Umgang mit dem Computer von den Schüler*innen benötigt. Nach Beendigung der Studie erhält jede teilnehmende Deutschlehrkraft ein differenziertes Feedback zu den Lernständen ihrer einzelnen Schüler*innen im Bereich Rechtschreibkompetenz. Während der Studie werden die Ergebnisse und die mit dem ReKoMe erhobenen Lernverläufe zur Rechtschreibkompetenz der Schüler*innen den Lehrkräften nicht bekannt gegeben, um einen Übungseffekt für das ReKoMe zu verhindern. Zur Überprüfung der externen Validität des ReKoMe wird zum fünften Messzeit-

punkt die Rechtschreibkompetenz der Schüler*innen mit der Hamburger Schreib-Probe (HSP) erhoben und mittels eines Lehrerfragebogens auf einer Skala von eins bis sechs in Schulnoten eingeschätzt. Im Zuge der Änderungen der Datenschutz-Grundverordnung 2018 musste die interne Codierung der Zugangspasswörter der Onlineplattform Levumi verändert werden. Dadurch gingen Daten verloren. Es nicht mehr möglich, die individuellen Ergebnisse der HSP und der eingeschätzten Rechtschreibleistungen durch die Lehrkräfte mit den Ergebnissen des ReKoMe zu vergleichen.

9.3 Datenanalysen

Die statistischen Analysen werden im Folgenden für die Testungen zum dritten und vierten Messzeitpunkt durchgeführt. Ab dem dritten Zeitpunkt hatte sich eine Routine bei der Durchführung der Testung eingestellt und die Schüler*innen wurden im Umgang mit dem Computer sicherer. Zum fünften Messzeitpunkt herrschte eine „Testmüdigkeit“bei den Schüler*innen, die sich durch den hohen organisatorischen Aufwand durch die genannten Rahmenbedingungen (vgl. Kap. 9.2) und die anstehenden Schulferien in der Woche der Testung erklären lässt. Basis für die folgenden Analysen bildet eine Stichprobe zum dritten Messzeitpunkt von insgesamt 144 Schüler*innen und zum vierten Messzeitpunkt von 123 Schüler*innen im durchschnittlichen Alter von 8.8 Jahren. Für die Validierung von ReKoMe liegen zum dritten Messzeitpunkt insgesamt 4926 und zum vierten Messzeitpunkt 5344 beantwortete Testitems vor (vgl. Tab. 9.1). Zu beiden Messzeitpunkten sind mehr weibliche Schüler*innen in der Stichprobe vertreten (3. MZP = 56% weiblich; 4. MZP = 54% weiblich). Einen diagnostizierten Förderbedarf im Bereich Deutsch haben 11,4% der Schüler*innen zum dritten Messzeitpunkt und 12% der Schüler*innen der Stichprobe zum vierten Messzeitpunkt. In der Stichprobe des dritten Messzeitpunkts ist der Anteil der Schüler*innen mit einem Migrationshintergrund mit 23% im Vergleich zu der Stichprobe zum vierten Messzeitpunkt mit 16% höher.

Zur Vorbereitung der Durchführung der Datenanalysen ist eine Bereinigung und Aufarbeitung des Datensatzes notwendig. Es müssen Antwortmuster und -verhalten, die auf „Spaßantworten“ oder „Weiterklicken“ beruhen, gelöscht werden. Um die auf der Onlinelernplattform Levumi registrierten Antworten weiter zu analysieren, ist eine aufwendige Aufbereitung notwendig. Sobald dies geschehen ist, können die Ergebnisse von der Plattform in Excel exportiert werden. Die exportierten Ergebnisse enthalten die jeweiligen Schreiblösungen und jeweilige schriftliche Informationen (string Variablen) zur Ergebnisanalyse. Durch die Programmierung eines Codes im Programm R können die schriftlichen Informationen (string Variablen) in numerische Informationen (numeric Variablen) umgewandelt werden. Dies stellt die Grundvoraussetzung für die statistischen Analysen dar. Während der Testung konnten maximal 53 Wörter von den Schüler*innen innerhalb der Testbearbeitungszeit von 20 Minuten geschrieben werden. Nicht alle Testwörter wurden dabei von allen Schüler*innen innerhalb der Testzeit bearbeitet. Zudem wurden durch einige Schüler*innen das Testwort nicht bearbeitet, sondern entweder nur weitergeglickt oder eine Spaßantwort gegeben (z.B. „Popo“). Um diese Fälle wurde der Datensatz be-

reinigt. In Tabelle 9.1 ist die bereinigte Datengrundlage [1] für die jeweiligen Skalen und unterschiedlichen Messzeitpunkten dargestellt.

Tabelle 9.1: Datengrundlage

Skala	MZP	N	Lup	n
Quan	3	144	52	1784
	4	123	52	2040
PhonSilb	3	146	29	1492
	4	123	29	1521
Morph	3	142	24	1043
	4	122	24	1101
Peri	3	132	12	438
	4	115	12	509
Wortbil	3	67	4	169
	4	65	4	173

Anmerkungen. N= Anzahl der Schüler*innen, die Items der jeweiligen Skala beantwortet haben. Lup = Anzahl der Lupenstellen, die in den jeweiligen Skalen enthalten sind. n = Anzahl der Testergebnisse, die pro Skala vorliegen und die Grundlage für die statistischen Berechnungen sind.

Gemäß der vorgeschlagenen Standardanalysen zur Prüfung der Gütekriterien eines Tests zur Lernverlaufsdiagnostik nach Wilbert & Linnemann (2011) werden folgende statistische Analysen auf Basis der Klassischen Testtheorie und der Item-Response-Theorie [2] mit dem Programm GNU R vorgenommen (vgl. Kap. 7.2):

Deskriptive Analysen: Das ReKoMe besteht aus insgesamt fünf Skalen, die unterschiedliche Kompetenzen der Rechtschreibkompetenz messen. Eine erste Beurteilung der Items der jeweiligen Skalen des ReKoMe erfolgt anhand von Itemanalysen auf Basis der klassischen Test Theorie (KTT) unter der Verwendung des Pakets *p*sych. Inwiefern Unterschiede der durchschnittlichen Itemschwierigkeiten auf Lernzuwächse zurückzuführen sind, wird mit einem t-Test geprüft. Die Berechnung der Retest- und Parallel-Reliabilität basiert auf den empirischen Korrelationen zwischen den Testwerten paralleler Tests (Schermelleh-Engel & Gäde, 2020). Die Testaufgaben des ReKoMe sind zu den jeweiligen Testungen nicht parallel. Als Äquivalent zur Retest-Reliabilität wird deshalb ein korrelativer Zusammenhang zwischen zwei aufeinanderfolgenden Messzeitpunkten berechnet.

Dimensionalitätsprüfung: Inwiefern sich das dem Test ReKoMe zugrunde gelegte theoretische Konstrukt bzw. die einzelnen Teilkompetenzen [3] des sprachsystematischen Rechtschreibkompetenzmodells in den vorliegenden Daten empirisch

[1] Die Ergebnisse des Shapiro-Wilk-Test zeigen, dass die Skalen nicht normalverteilt (p>.05) sind.
[2] Die Ergebnisse der Analysen aller Items ist im Anhang dargestellt (vgl. Kap. 11.2).
[3] *P*honologisch-Silbisches Prinzip, Morphologisches Prinzip, Wortbilungs Prinzip, Peripheriebereich

belegen lassen, wird mit dem Vergleich eines einfaktoriellen und eines vierfaktoriellen Modells mittels einer konfirmatorischen Faktorenanalyse geprüft. Die Ergebnisse werden durch die deskreptiven Gütemaße Akaike Information Criterion (AIC) und Bayesian Information Criterion (BIC) bewertet.

Schätzung der Modellparameter: Die Schätzung der Itemparameter für das Raschmodell wird mittels des R-Pakets *e*Rm unter der Anwendung der Conditional ML-Methode (CML) in einem ersten Schritt durchgeführt und anhand der Infit-Werte beurteilt. Diese sind im Gegensatz zu den Outfit-Werten im mittleren Leistungsniveau zuverlässiger und weniger anfällig für Ausreißer in den Werten (Linacre et al., 2002). Items, die keine zufriedenstellende Passung zum Rasch-Modell aufweisen ($.75 \leq$ Infit/Outfit≤ 1.3), werden vor den Berechnungen ausgeschlossen.

Vergleich Item- und Personenparameter: Personenparameter und Itemparameter können mittels Raschanalysen auf der gleichen Skala abgebildet werden. Dies erlaubt eine Beurteilung der Testaufgaben bezüglich der Angemessenheit der Schwierigkeitsgrade. Die Verteilung der Personenparameter und die Position der Items auf der latenten Dimension werden anhand einer Person-Item-Map dargestellt Die Beurteilung der übergeordneten Messgenauigkeit der Skalen wird mit dem Paket Tam berechnet und als EAP Reliabilität und WLE Reliabilität berichtet.

Itemhomogenität und Testfairness: Der bedingte Likelihood-Quotienten-Test (Andersen, 1973) wird gerechnet, um zu prüfen, ob sich die Itemparameter zwischen zwei Gruppen (Teilungskriterium: Median, Geschlecht) unterscheiden und eine Itemhomogenität vorliegt.

Modellvergleich: Inwiefern ein 2PL-Modell mit frei schätzbaren Ladungen (Birnbaum-Modell) die Daten signifikant besser erklärt als ein 1PL-Modell mit identischen Ladungen (Rasch-Modell), wird anhand eines Likelihood-Ratio-Tests geprüft. Dazu werden die Modellparameter berechnet und anschließend mit einer ANOVA ermittelt, welches Modell die Daten besser erklärt. Für die Schätzung der Parameter des zweiparametrigen logistischen Modells wird das Paket *l*tm verwendet. Als deskriptives Gütekriterium wird das *B*ayesian Information Criterion (BIC) herangezogen, da es die Parameteranzahl und Stichprobengröße berücksichtigt (Rost, 2004). Im Vergleich von zwei Modellen zeigt ein kleinerer BIC-Wert, dass das Verhältnis von Modellfit und Modellsparsamkeit besser passt als das Vergleichsmodell (Gäde et al., 2020).

9.4 Ergebnisse - Validierung der Skala Quan

Die Skala Quan misst die Fähigkeit, Wörter auf Ganzwortebene orthographisch richtig schreiben zu können.

9.4.1 Itemanalysen auf Basis der klassischen Testtheorie

Die Ergebnisse [4] der Itemanalysen auf Basis der Klassischen Testtheorie für die Skala Quan zeigen (vgl. Tab. 9.2), dass sich die Items sehr gut zur Abbildung der verschiedenen Merkmalsausprägungen der Testpersonen der Stichprobe eignen. Die durchschnittliche Itemschwierigkeit liegt zum dritten Messzeitpunkt im mittleren Schwierigkeitsbereich, es wurden durchschnittlich 53% der Items richtig gelöst. Zum vierten Messzeitpunkt wurden durchschnittlich 8% mehr Items richtig gelöst. Die Ergebnisse eines t-Tests zeigen, dass der Lernzuwachs signifikant ($p < .01$) ist. Die Items differenzieren gut ($r_{it3} = .42$; $r_{it4} = .43$) zwischen Personen mit höheren und Personen mit niedrigeren Merkmalsausprägungen. Mit einer sehr hohen internen Konsistenz ($\alpha_{3+4} = .93$; $r_{tt3+4} = .96$) misst die Skala Quan zu Messzeitpunkt drei und vier sehr gut die Fähigkeit, Wörter auf Ganzwortebene orthografisch richtig schreiben zu können.

Tabelle 9.2: Itemanalysen KTT- Skala Quan

Zeitpunkt	Itemschwierigkeit		Trennschärfe		interne Konsistenz	
	$M(SD)$	Min-Max	$M(SD)$	Min-Max	α	Split-Half
MZP3	.53 (.24)	.03-.96	.42 (.13)	.03-.64	.93	.96
MZP4	.58 (.25)	.02-.97	.43 (.17)	.05-.72	.93	.96

9.4.1.1 Retestreliabilität

Die Berechnung der Retest- und Parallel-Reliabilität basiert auf den empirischen Korrelationen zwischen den Testwerten paralleler Tests (Schermelleh-Engel & Gäde, 2020). Es wurden Korrelationen zwischen aufeinanderfolgenden Testzeitpunkten als Äquivalent zur Retest-Reliabilität berechnet. Diese liegen für die Skalen PhonSilb, Morph und Peri mit einem r zwischen .93 - .99 ($p < .01$) in einem sehr hohen Bereich. Zwischen den aufeinanderfolgenden Testungen der Skala Wortbil besteht kein korrelativer Zusammenhang ($p < .05$).

9.4.1.2 Dimensionalitätsprüfung

Im sprachsystematischen Rechtschreibkompetenzmodell wird eine mehrfaktorielle Struktur postuliert. Inwiefern sich das dem ReKoMe zugrunde gelegte theoretische Konstrukt bzw. die einzelnen Teilkompetenzen des sprachsystematischen Rechtschreibkompetenzmodells in den vorliegenden Daten empirisch belegen lassen, wird mit dem Vergleich eines einfaktoriellen und eines vierfaktoriellen Modells mittels einer konfirmatorischen Faktorenanalyse hinsichtlich ihrer Güte geprüft. Basis der Analysen bilden die Daten der Skala Quan zum vierten Messzeitpunkt (N = 123). Im Vergleich der zwei Modelle anhand des

[4] In der Skala sind insgesamt 53 Items enthalten. Für die Durchführbarkeit der Analysen musste ein Item von den Berechnungen ausgeschlossen werden.

Akaike Information Criterion (AIC) und Bayesian Information Criterion (BIC) als deskriptives Gütemaß weist das vierfaktorielle Modell mit kleineren Werten bezüglich des AIC und BIC auf eine bessere Passung zwischen dem theoretisch zugrunde gelegten Konstrukt und der Daten hin. Dieses Ergebnis steht im Einklang mit dem Rahmenmodell des sprachsystematischen Rechtschreibkompetenzmodells, das die hier überprüften und bestätigten vier Faktoren postuliert (vgl. Tab. 9.3).

Tabelle 9.3: Modellgeltungstest - konfirmatorischen Faktorenanalyse

Dim	x2	df	x2/df	RMSEA	SRMR	CFI	TLI	AIC	BIC
1-Faktor	1246.57	740	1.68	.08 [.07;.08]	.10	.64	.62	4675.24	5015.59
4-Faktor	1209.30	734	1.65	.07 [.07;.08]	.09	.66	.64	4637.90	4995.28

*Anmerkungen.*1-Faktor = einfaktorielles Modell; 4-Faktor = vierfaktorielles Modell.

9.4.2 Itemanalysen auf Basis der Item-Response-Theorie

9.4.2.1 Schätzung der Modellparameter

Die Ergebnisse der Skalierung der Items am eindimensionalen Rasch-Modell zum dritten und vierten Messzeitpunkt (vgl. Tab. 9.4) zeigen, dass mit einem durchschnittlichen InfitMNSQ von .98 und OutfitMNSQ zwischen 1.04 und 1.08 eine sehr gute Passung der Items zum Rasch-Modell ($.75 \leq$ Infit/Outfit ≤ 1.3) bestätigt werden kann . Die Werte liegen sehr nahe am Erwartungswert von 1 (Bond et al., 2020). Die Items differenzieren gut ($r_{it3} = .42$; $r_{it4} = .44$) zwischen Personen mit höheren und Personen mit niedrigeren Merkmalsausprägungen.

Tabelle 9.4: Itemanalysen IRT - Skala Quan

Zeitpunkt	Itemschwierigkeit		Trennschärfe		InfitMNSQ		OutfitMNSQ	
	$M(SD)$	Min-Max	$M(SD)$	Min-Max	$M(SD)$	Min-Max	$M(SD)$	Min-Max
MZP3	0 (1.76)	-3.68-5.57	.42 (.13)	.03-.64	.98 (.13)	.76-1.3	1.08 (.60)	.21-4.04
MZP4	0 (1.69)	-3.33-4.26	.44 (.15)	.09-.65	.98 (.16)	.75-1.42	1.04 (.59)	.27-3.19

Zum dritten Messzeitpunkt weisen zwei und zum vierten Messzeitpunkt fünf Items keine ausreichende Passung (InfitMNSQ $< .75$) zum Rasch-Modell auf und wurden vor den Berechnungen ausgeschlossen. Das Ausgangsmodell des dritten Messzeitpunkts besteht insgesamt aus 51 Items und das Ausgangsmodell des vierten Messzeitpunkts aus 48 Items. Im Itempool befinden sich zu beiden Messzeitpunkten auch Items, dessen Outfit-Werte nicht im Bereich von $.75 \leq$ Outfit ≤ 1.3 liegen. Da zu beiden Zeitpunkten die Werte des OutfitMNSQ-Werts im Mittel jedoch sehr nahe am Erwartungswert von 1 verortet sind,

ist davon auszugehen, dass es sich um kein systematisches Problem der Skala handelt, sondern vielmehr dem Datensatz und der Erhebungssituation geschuldet ist (vgl. Kap. 9.3). Die Items werden vor den Analysen nicht ausgeschlossen.

9.4.2.2 Vergleich der Item- und Personenparameter

Die Items der Skala Quan erfassen die unterschiedlichen Personenfähigkeiten der Schüler*innen gut (EAP Rel.$_3$: .89; WLE Rel.$_3$: .88; EAP Rel.$_4$: .89; WLE Rel.$_4$: .88). Der Vergleich der Item- und Personenparameter zeigt (vgl. Abb. 9.1; 9.2), dass die Items über den gesamten Bereich der Personenfähigkeiten streuen. Zum dritten Messzeitpunkt (vgl. Abb. 9.1) liegen zwei sehr leichte Items und ein schweres Item außerhalb der Personenfähigkeiten. Im Vergleich zum dritten Messzeitpunkt fehlen zum vierten Messzeitpunkt (vgl. Abb. 9.2) Items für sehr leistungsstarke Schüler*innen.

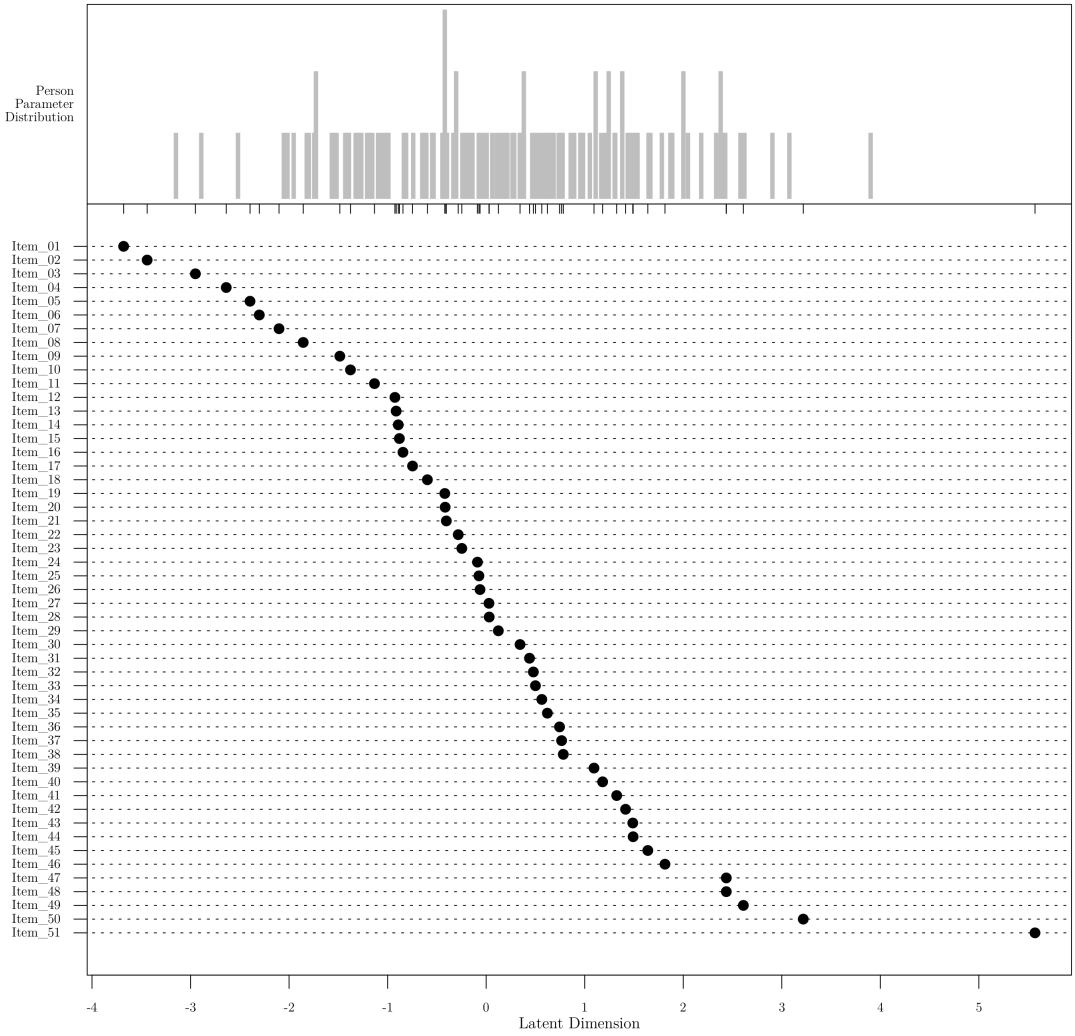

Abbildung 9.1: Person-Item Map. Verteilung der Personenparameter und Position der Items auf latenter Dimension. Skala Quan; MZP3. LA = 51.

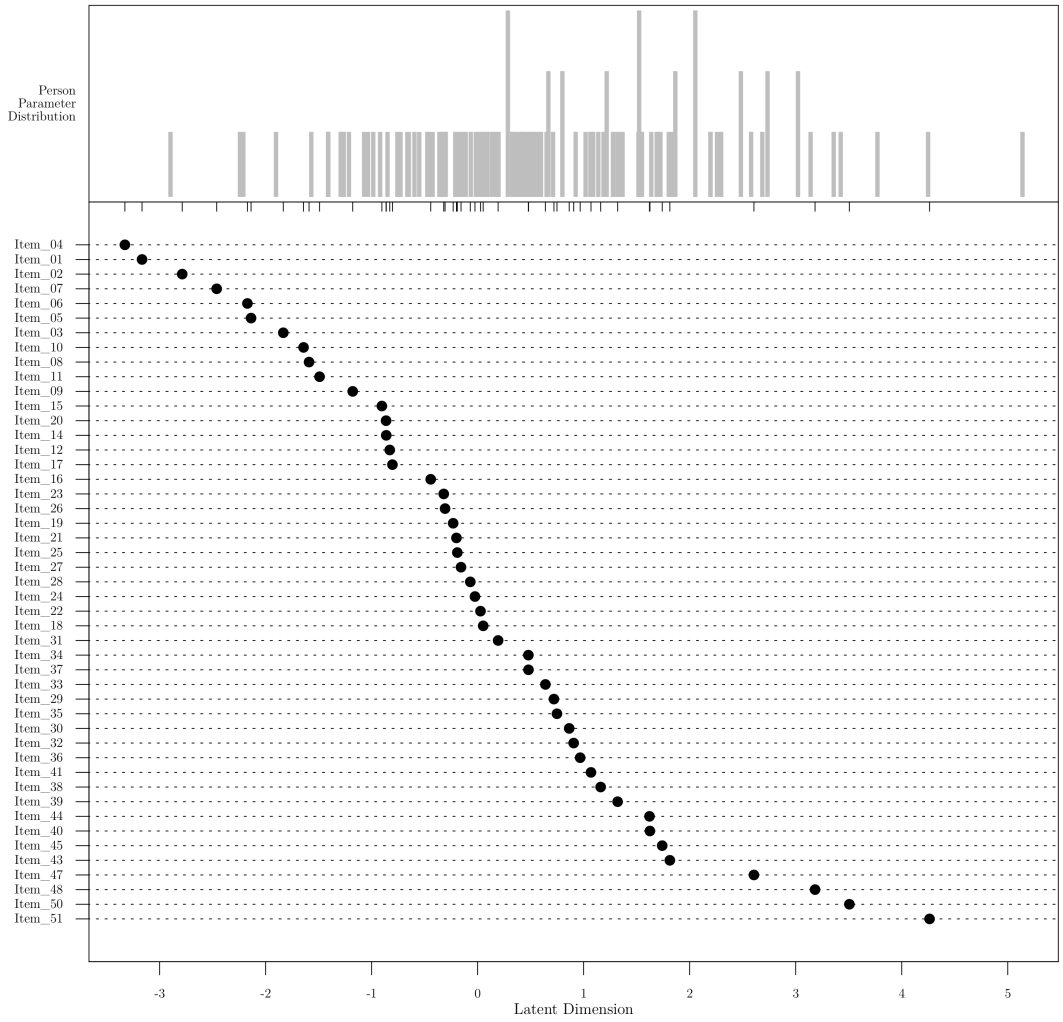

Abbildung 9.2: Person-Item Map. Verteilung der Personenparameter und Position der Items auf latenter Dimension. Skala Quan; MZP4. LA = 48.

9.4.2.3 Prüfung der Itemhomogenität

Die Ergebnisse des Likelihood-Quotienten-Tests der Skala Quan zeigen, dass sich die Schätzungen der Item-Parameter [5] nicht signifikant aufgeteilt nach dem Median (vgl. Tab. 9.5) unterscheiden und keine Modellverletzungen ($p < .01$) vorliegen (vgl. Tab. 9.5, 9.3, 9.4).

[5] Einzelne Items weisen keine passende Antwortmuster zur Durchführung des Likelihood-Quotienten-Tests auf und werden automatisch von den Analysen ausgeschlossen. Der Ausgangsitempool zur Prüfung der Itemhomogenität besteht zum dritten Messzeitpunkt aus 39 Items und zum vierten Messzeitpunkt aus 33 Items.

Tabelle 9.5: Itemhomogenität - Skala Quan

	LU	LA	*LR*	*df*	*p*
MZP3	53	39	21.56	38	.99
MZP4	53	33	41.90	32	.11

Anmerkungen. Ergebnisse des Likelihood-Quotienten-Tests für Skala Quan; Teilungskriterium Median; Bonferroni-Korrektur ($p - Wert < .01$).
LU = Anzahl der Lupenstellen im Ursprungsitempool; LA = Anzahl der Lupenstellen, die auf Itemhomogenität geprüft werden konnten.

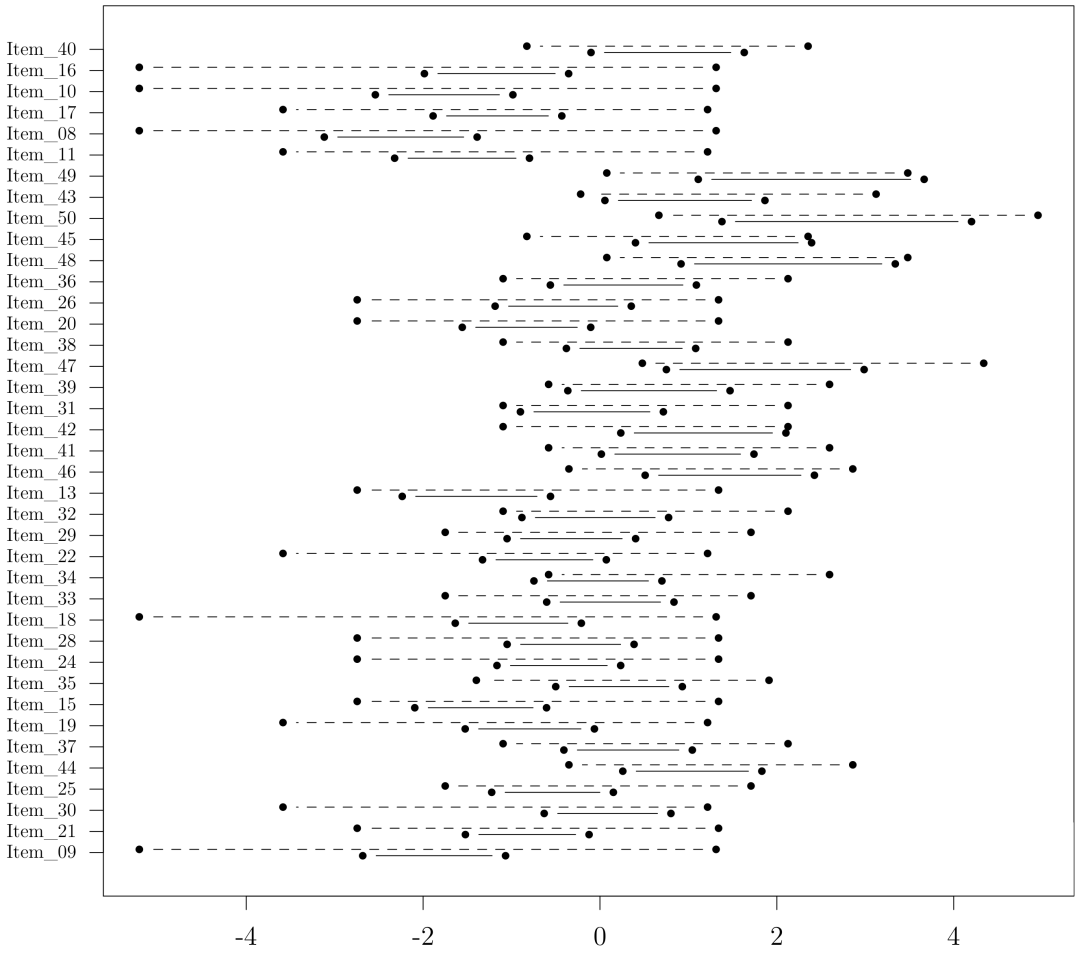

Abbildung 9.3: Plot der Konfidenzintervalle. Itemschwierigkeiten aufgeteilt nach Teilungskriterium Median; Bonferroni-Korrektur. Skala Quan; MZP3.

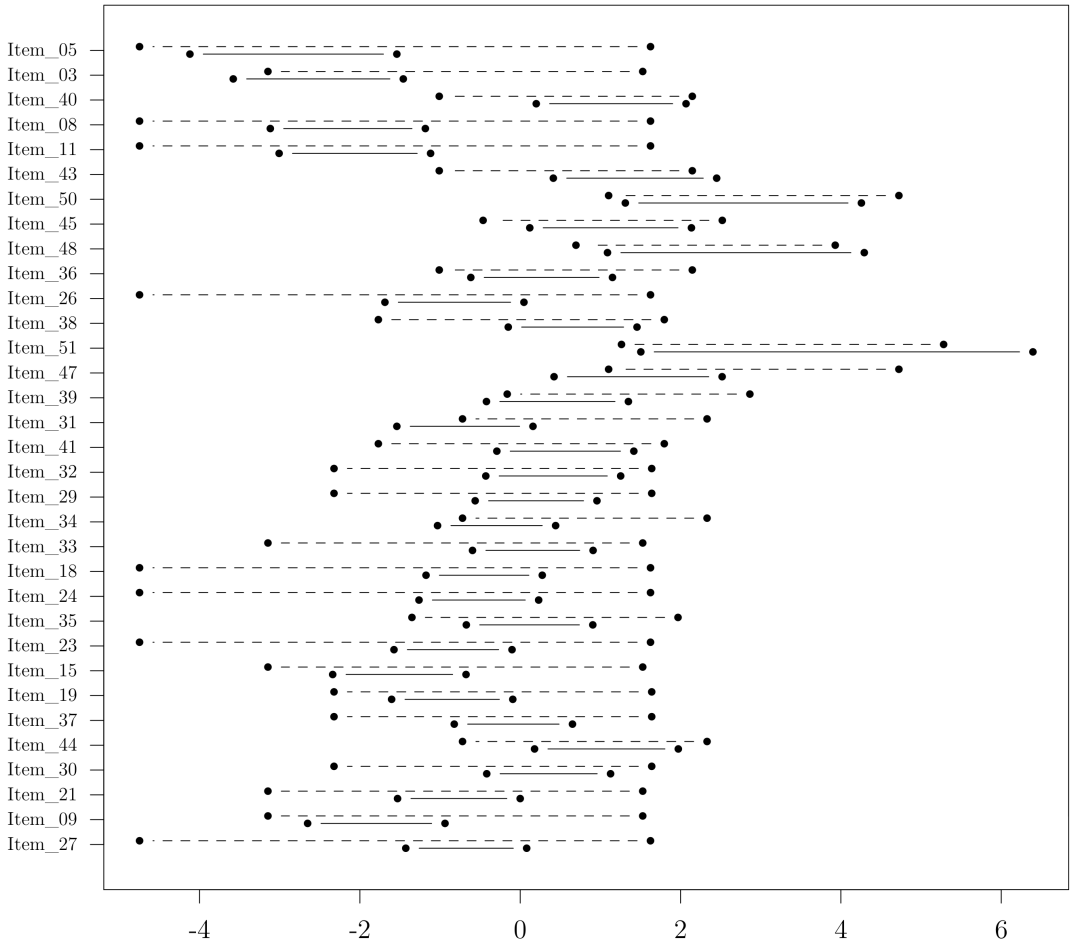

Abbildung 9.4: Plot der Konfidenzintervalle. Itemschwierigkeiten aufgeteilt nach Teilungskriterium Median; Bonferroni-Korrektur. Skala Quan; MZP4.

9.4.2.4 Prüfung der Testfairness

Die Ergebnisse des Likelihood-Quotienten-Tests zur Prüfung der Testfairness (vgl. Tab. 9.6) der Skala Quan [6] zeigen, dass sich die Schätzungen der Item-Parameter nicht signifikant aufgeteilt nach dem Geschlecht unterscheiden und kein Differential Item Functioning (DIF) ($p < .01$) vorliegt. Schüler*innen mit gleichem Fähigkeitsniveau erreichen die gleichen Personenscores im ReKoMe (vgl. Tab. 9.6, 9.5, 9.6).

[6] Einzelne Items weisen keine passende Antwortmuster zur Durchführung des Likelihood-Quotienten-Tests auf und werden automatisch von den Analysen ausgeschlossen. Der Ausgangsitempool zur Prüfung der Testfairness besteht zum dritten Messzeitpunkt aus 50 Items und zum vierten Messzeitpunkt aus 47 Items.

Tabelle 9.6: Testfairness - Skala Quan

	LU	LA	LR	*df*	*p*
MZP3	53	50	52.95	49	.32
MZP4	53	47	28.27	46	.98

Anmerkungen. Ergebnisse des Likelihood-Quotienten-Tests für Skala Quan.; Teilungskriterium Geschlecht; Bonferroni-Korrektur ($p - Wert < .01$).
LU = Anzahl der Lupenstellen im Ursprungsitempool; LA = Anzahl der Lupenstellen, die auf Itemhomogenität geprüft werden konnten.

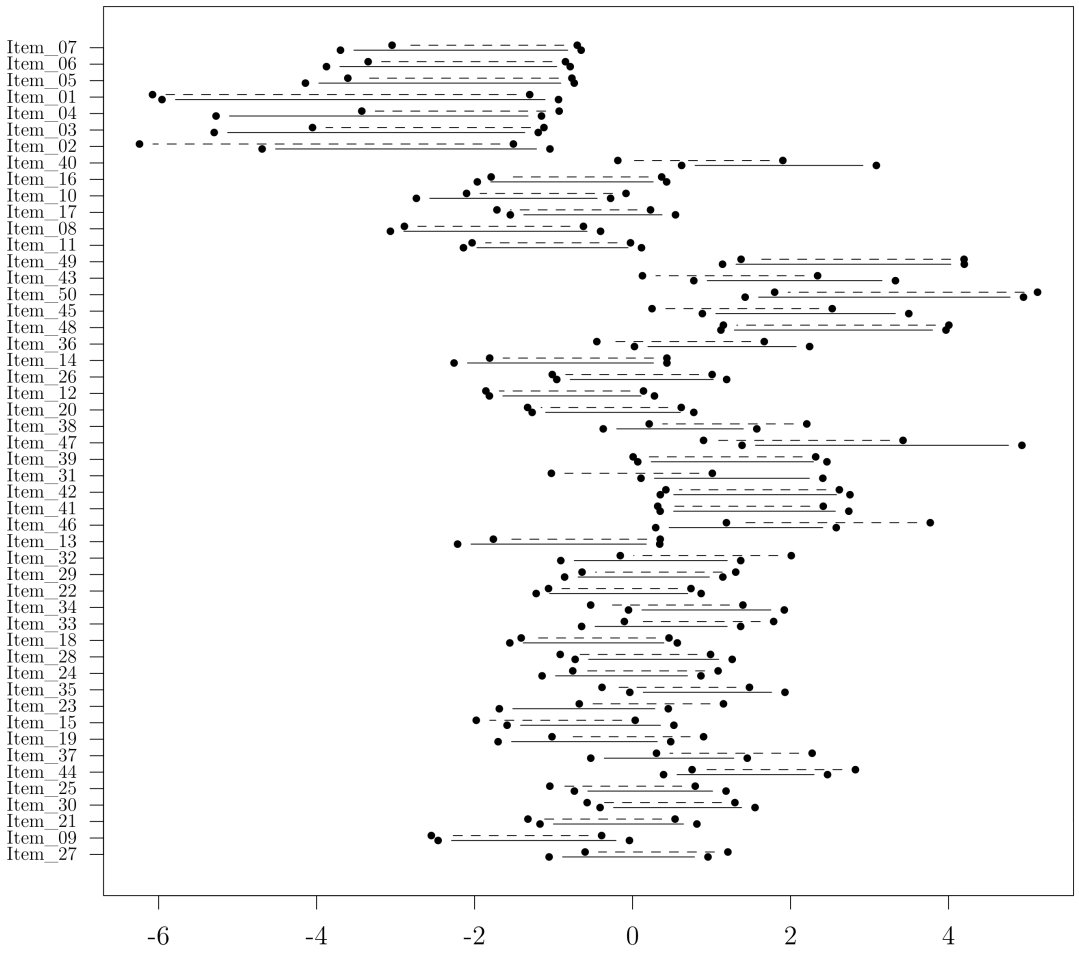

Abbildung 9.5: DIF-Plot. Itemschwierigkeiten aufgeteilt nach Geschlecht mit korrigierten Konfidenzintervall; Bonferroni-Korrektur. Skala Quan; MZP 3.

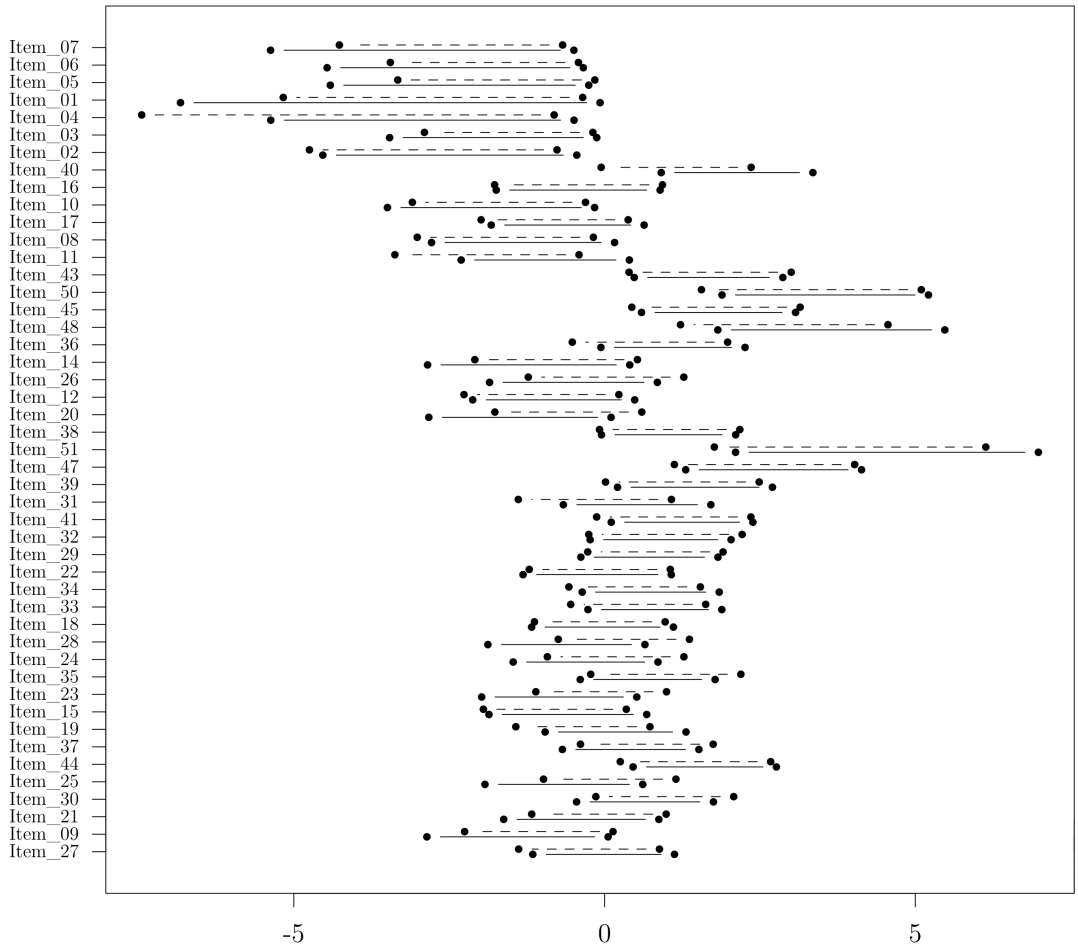

Abbildung 9.6: DIF-Plot. Itemschwierigkeiten aufgeteilt nach Geschlecht mit korrigierten Konfidenzintervall; Bonferroni-Korrektur. Skala Quan; MZP 4.

9.4.2.5 Modellvergleich

Der Modellvergleich der Skala Quan zwischen dem Rasch-Modell und zwischen dem Birnbaum-Modell bestätigt, dass die Daten besser durch das Rasch-Modell ($p<.001$) erklärt werden. Der BIC-Wert zeigt (vgl. Tab. 9.7) jeweils eine bessere Datenbeschreibung mit einem niedrigeren Wert an.

Tabelle 9.7: Modellvergleich - Skala Quan

Modell	MZP3		MZP4	
	AIC	BIC	AIC	BIC
1 PL	5824.98	5979.41	4595.92	4730.52
2 PL	5805.61	6108.53	4556.78	4820.35

Anmerkungen. 1 PL = Rasch-Modell; 2 PL = Birnbaum-Modell.

9.5 Ergebnisse - Validierung der Skala PhonSilb

Die Skala PhonSilb misst die Fähigkeit, einen Bezug zwischen Schriftstruktur und Lautstruktur unter Berücksichtigung der silbenstrukturellen Informationen (Silbeanfangsrand, Silbenendrand und Silbenschnitt) herstellen zu können.

9.5.1 Itemanalysen auf Basis der klassischen Testtheorie

Die Ergebnisse der Itemanalysen auf Basis der Klassischen Testtheorie zeigen (vgl. Tab. 9.8), dass die durchschnittliche Itemschwierigkeit der Skala PhonSilb eher leicht war. Zum dritten Messzeitpunkt wurden durchschnittlich 66% und zum vierten Messzeitpunkt durchschnittlich 70% der Items richtig gelöst. Die Ergebnisse eines t-Tests bestätigt, dass der Lernzuwachs zwischen den Messzeitpunkten signifikant ($p < .05$) ist. Die Items differenzieren nicht im ausreichenden Maße ($r_{it3+4} = .34$) zwischen Personen mit höheren und Personen mit niedrigeren Merkmalsausprägungen. Die interne Konsistenz der Skala PhonSilb ($\alpha_3 = .83$, $\alpha_4 = .84$; $r_{tt3} = .88$, $r_{tt3} = .90$) ist hoch.

Tabelle 9.8: Itemanalysen KTT - Skala PhonSilb

Zeitpunkt	Itemschwierigkeit		Trennschärfe		interne Konsistenz	
	$M(SD)$	Min-Max	$M(SD)$	Min-Max	α	Split-Half
MZP3	.66 (.27)	.06-.95	.34 (.15)	-.01-.67	.83	.88
MZP4	.70 (.26)	.07-.98	.34 (.18)	-.01-.67	.84	.90

9.5.2 Itemanalysen auf Basis der Item-Response-Theorie

9.5.2.1 Schätzung der Modellparameter

Die Ergebnisse der Skalierung der Items am eindimensionalen Rasch-Modell zum dritten und vierten Messzeitpunkt (vgl. Tab. 9.9) zeigen, dass mit einem durchschnittlichen InfitMNSQ$_3$ von .94 und OutfitMNSQ$_4$ zwischen .97 und 1.14 eine sehr gute Passung der Items zum Rasch-Modell bestätigt werden kann ($.75 \leq$ Infit/Outfit ≤ 1.3). Die Werte liegen sehr nahe am Erwartungswert von 1 (Bond et al., 2020).
Die Items differenzieren gut ($r_{it} = .38$) zwischen Personen mit höheren und Personen mit niedrigeren Merkmalsausprägungen. Der Ursprungsitempool der Skala PhonSilb beinhaltet 30 Items. Zum dritten Messzeitpunkt wurde ein Item vor den Berechnungen ausgeschlossen, da es nicht raschkonform (Infit $< .75$) ist. Ein zweites Item weist keine passende Datenstruktur für die Analysen auf, es wurde immer richtig beantwortet und wurde automatisch von den Analysen ausgeschlossen. Im Itempool befinden sich zu beiden Messzeitpunkten auch Items, dessen Outfit-Werte (z.B. MZP 3: Min-Max = .47-3.42) nicht im Bereich von ($.75 \leq$ Outfit ≤ 1.3) liegen. Die gemittelten Werte des OutfitMNSQ-Werts

liegen zu beiden Zeitpunkten aber sehr nahe am Erwartungswert von 1 und wurden nicht von den Berechnungen ausgeschlossen, da von keinem systematischen Fehler der Skala ausgegangen wird. Der Ausgangsitempool zur Berechnung der Modellparameter besteht zum dritten Messzeitpunkt aus 28 Items und zum vierten Messzeitpunkt aus 29 Items.

Tabelle 9.9: Itemanalysen IRT- Skala PhonSilb

Zeitpunkt	Itemschwierigkeit		Trennschärfe		InfitMNSQ		OutfitMNSQ	
	$M(SD)$	Min-Max	$M(SD)$	Min-Max	$M(SD)$	Min-Max	$M(SD)$	Min-Max
MZP3	-.03 (1.9)	-2.67-4.43	.38 (.17)	-.07-.6	.94 (.15)	.65-1.25	1.14 (.64)	.47-3.42
MZP4	-.02 (2.0)	-3.36-4.69	.38 (.19)	.02-.73	.94 (.14)	.75-1.39	.97 (.54)	.39-3.24

9.5.2.2 Vergleich der Item- und Personenparameter

Die Items der Skala PhonSilb erfassen die unterschiedlichen Personenfähigkeiten der Schüler*innen gut (EAP Rel.$_3$: .76; WLE Rel.$_3$: .71; EAP Rel.$_4$: .79; WLE Rel.$_4$: .73). Der Vergleich der Item- und Personenparameter zeigt (vgl. Abb. 9.7; 9.8), dass die Items über den gesamten Bereich der Personenfähigkeiten streuen. Zum dritten Messzeitpunkt (vgl. Abb. 9.7) werden die Personenfähigkeiten der Schüler*innen durch die Items an den Randbereichen des Fähigkeitsniveaus nicht abgedeckt. Zum vierten Messzeitpunkt waren vier Items zu leicht. Es fehlen Items für sehr leistungsstarke Schüler*innen.

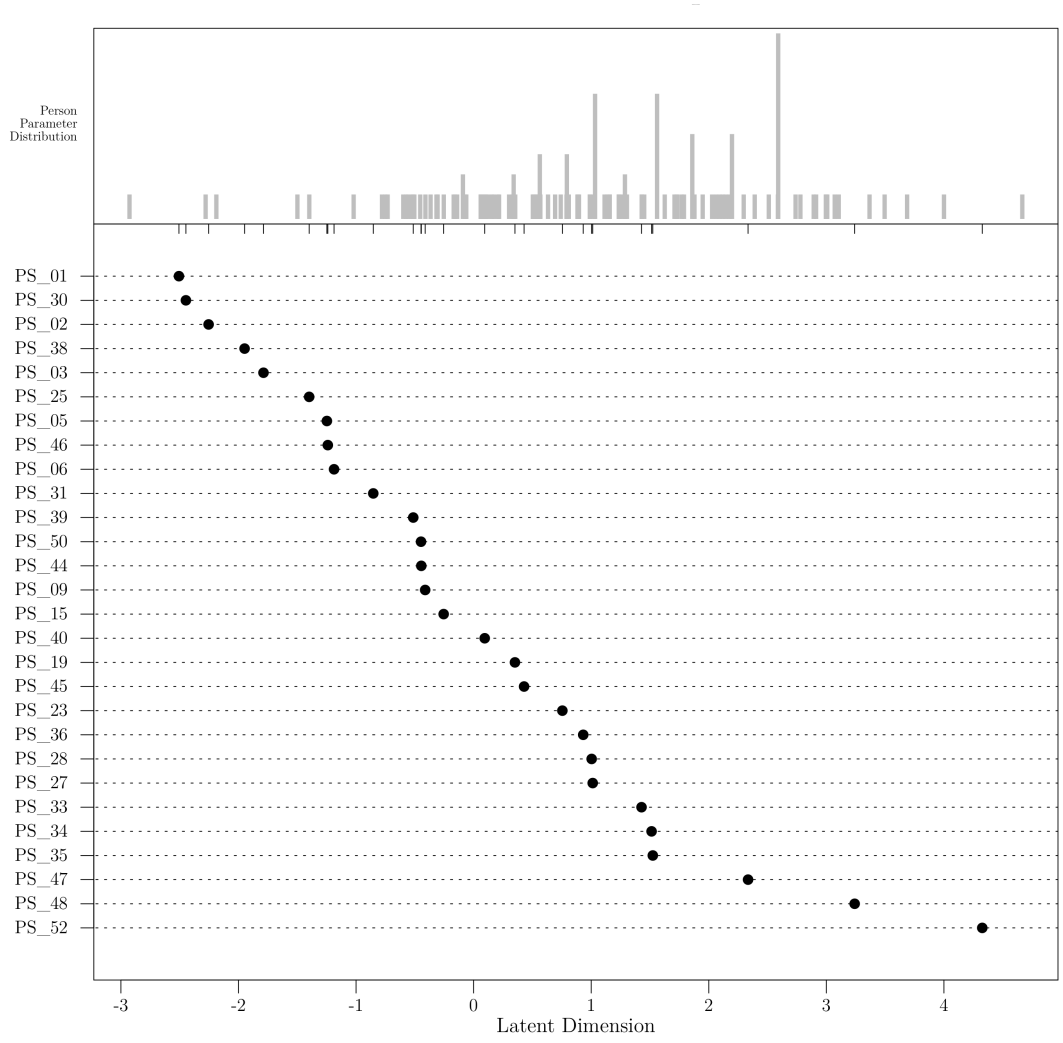

Abbildung 9.7: Person-Item Map. Verteilung der Personenparameter und Position der Items auf latenter Dimension. Skala PhonSilb; MZP3.

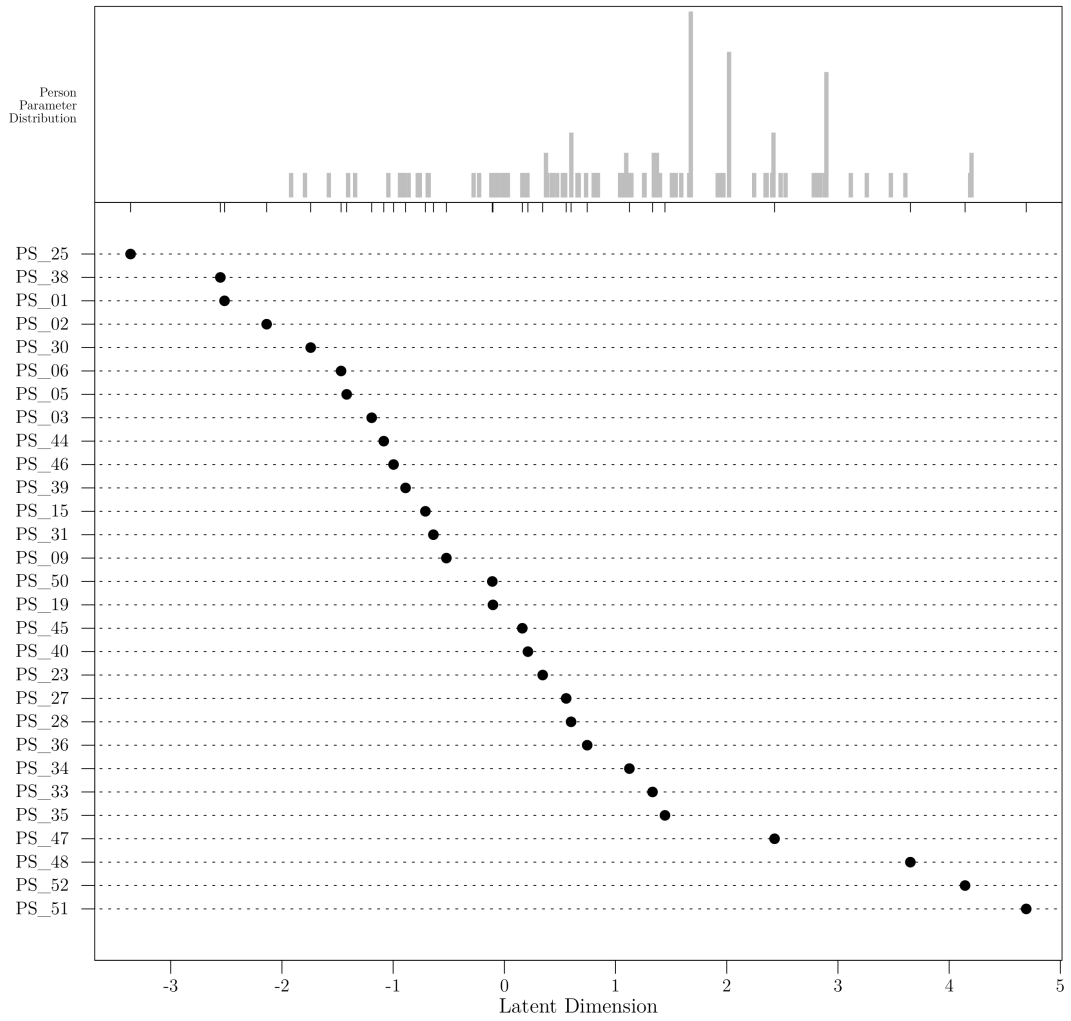

Abbildung 9.8: Person-Item Map. Verteilung der Personenparameter und Position der Items auf latenter Dimension. Skala PhonSilb; MZP4.

9.5.2.3 Prüfung der Itemhomogenität

Die Ergebnisse des Likelihood-Quotienten-Tests der Skala PhonSilb [7] zeigen, dass sich die Schätzungen der Item-Parameter nicht signifikant aufgeteilt nach dem Median unterscheiden und keine Modellverletzungen ($p < .01$) vorliegen (vgl. Tab. 9.10, Abb. 9.9, 9.10).

[7] Einzelne Items weisen keine passende Antwortmuster zur Durchführung des Likelihood-Quotienten-Tests auf und werden automatisch von den Analysen ausgeschlossen. Der Ausgangsitempool zur Prüfung der Testfairness besteht zum dritten Messzeitpunkt aus 21 Items und zum vierten Messzeitpunkt aus 23 Items.

Tabelle 9.10: Itemhomogenität - Skala PhonSilb

	LU	LA	LR	*df*	*p*
MZP3	30	21	27.31	20	.13
MZP4	30	23	26.06	22	.25

Anmerkungen. Ergebnisse des Likelihood-Quotienten-Tests für Skala PhonSilb; Teilungskriterium Median; Bonferroni-Korrektur ($p - Wert < .01$).
LU = Anzahl der Lupenstellen im Ursprungsitempool; LA = Anzahl der Lupenstellen, die auf Itemhomogenität geprüft werden konnten.

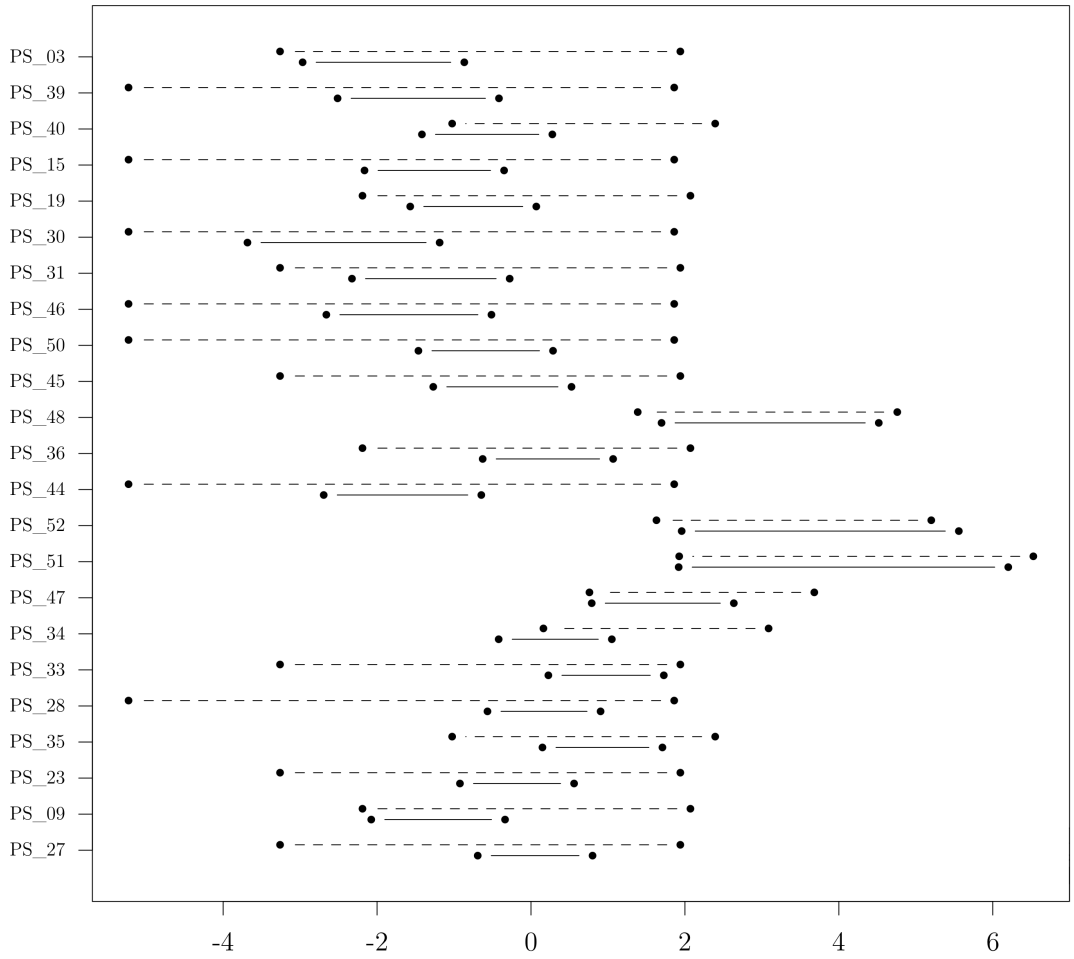

Abbildung 9.10: Plot der Konfidenzintervalle. Itemschwierigkeiten aufgeteilt nach Teilungskriterium Median; Bonferroni-Korrektur. Skala PhonSilb; MZP4.

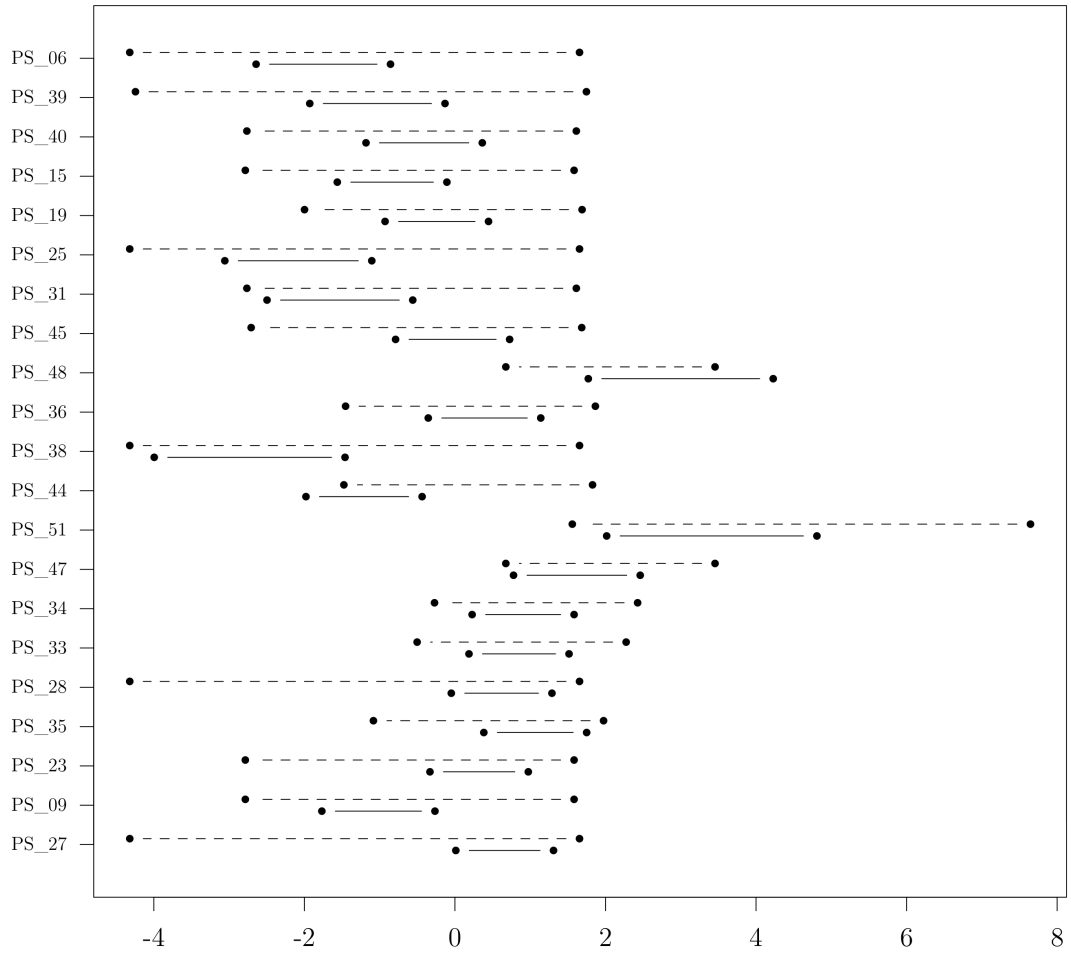

Abbildung 9.9: Plot der Konfidenzintervalle. Itemschwierigkeiten aufgeteilt nach Teilungskriterium Median; Bonferroni-Korrektur. Skala PhonSilb; MZP3.

9.5.2.4 Prüfung der Testfairness

Die Ergebnisse des Likelihood-Quotienten-Tests zur Prüfung der Testfairness (vgl. Tab. 9.11) der Skala PhonSilb [8] zeigen, dass sich die Schätzungen der Item-Parameter nicht signifikant aufgeteilt nach dem Geschlecht unterscheiden und kein Differential Item Functioning (DIF) ($p < .01$) vorliegt. Schüler*innen mit gleichem Fähigkeitsniveau erreichen die gleichen Personenscores im ReKoMe (vgl. Abb. 9.11, 9.12).

[8] Einzelne Items weisen keine passende Antwortmuster zur Durchführung des Likelihood-Quotienten-Tests auf und werden automatisch von den Analysen ausgeschlossen. Der Ausgangsitempool zur Prüfung der Testfairness besteht zum dritten Messzeitpunkt aus 23 Items und zum vierten Messzeitpunkt aus 25 Items.

Tabelle 9.11: Testfairness - Skala PhonSilb

	LU	LA	LR	*df*	*p*
MZP3	30	23	23.98	22	.35
MZP4	30	25	11.68	24	.98

Anmerkungen. Ergebnisse des Likelihood-Quotienten-Tests für Skala PhonSilb; Teilungskriterium Geschlecht; Bonferroni-Korrektur ($p - Wert < .01$).
LU = Anzahl der Lupenstellen im Ursprungsitempool; LA = Anzahl der Lupenstellen, die auf Itemhomogenität geprüft werden konnten.

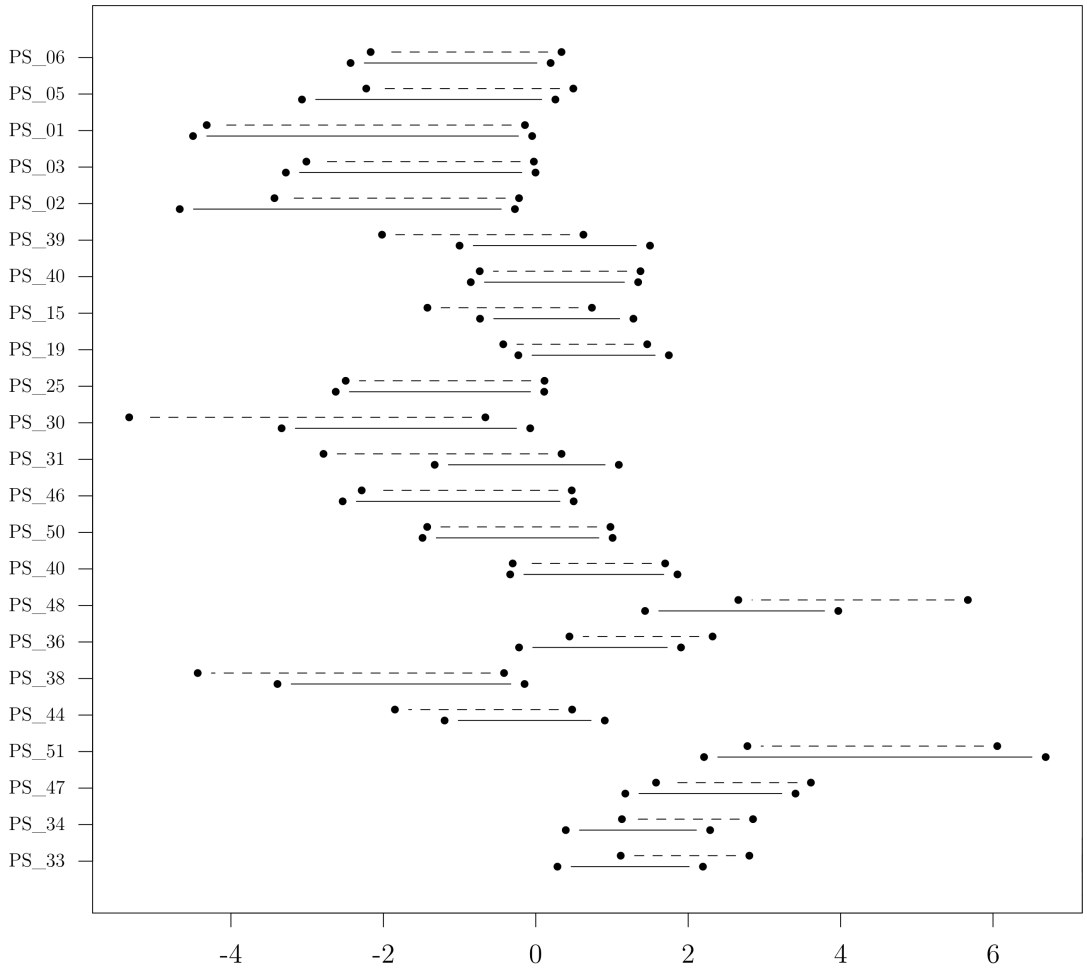

Abbildung 9.11: DIF-Plot. Itemschwierigkeiten aufgeteilt nach Geschlecht mit korrigierten Konfidenzintervall; Bonferroni-Korrektur. Skala PhonSilb; MZP 3.

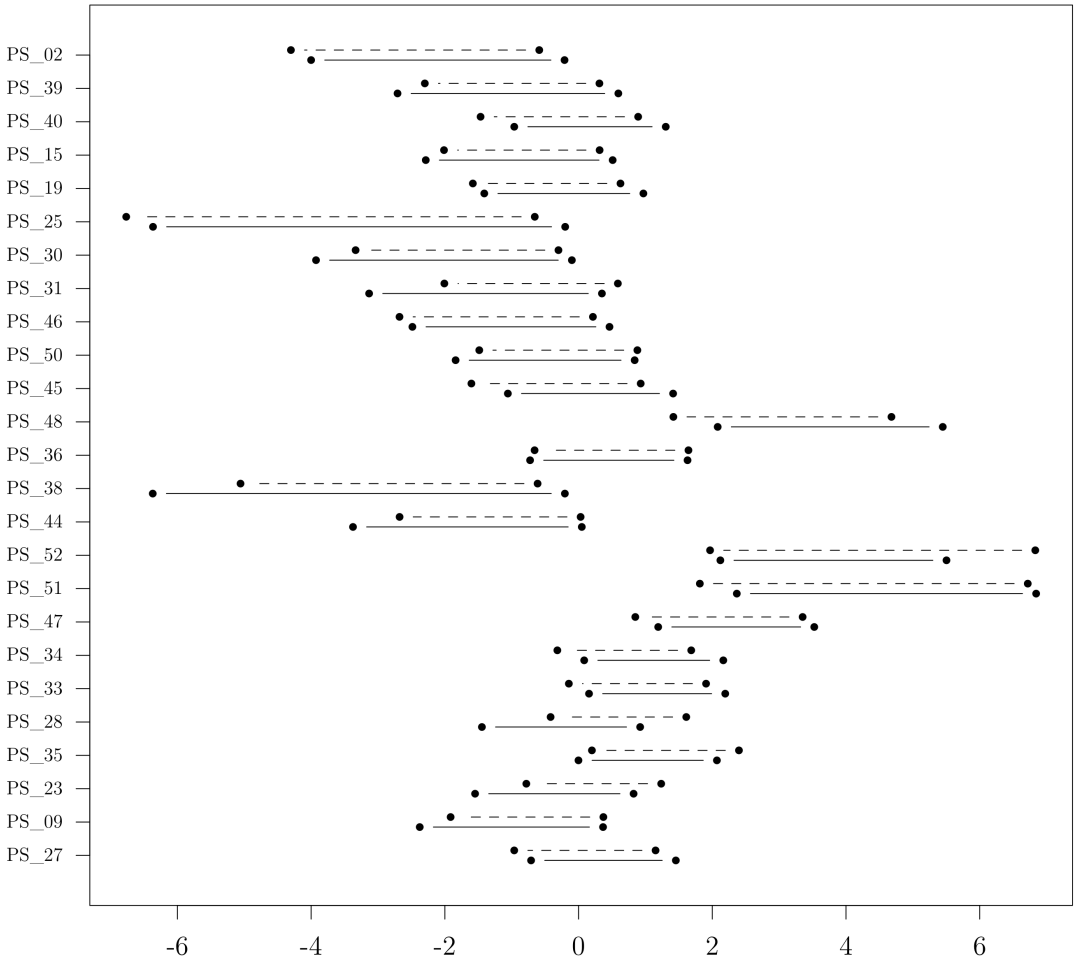

Abbildung 9.12: DIF-Plot. Itemschwierigkeiten aufgeteilt nach Geschlecht mit korrigierten Konfidenzintervall; Bonferroni-Korrektur. Skala PhonSilb; MZP 4.

9.5.2.5 Modellvergleich

Der Modellvergleich der Skala PhonSilb zwischen dem Rasch-Modell und dem Birnbaum-Modell bestätigt, dass die Daten besser durch das Rasch-Modell ($p < .001$) erklärt werden. Der AIC- und BIC-Wert zeigen jeweils eine bessere Datenbeschreibung mit einem niedrigeren Wert an.

Tabelle 9.12: Modellvergleich - Skala PhonSilb

Modell	MZP3		MZP4	
	AIC	BIC	AIC	BIC
1 PL	2981.39	3073.89	2415.62	2502.79
2 PL	2955.54	3134.56	2411.77	2580.50

Anmerkungen. 1 PL = Rasch-Modell; 2 PL = Birnbaum-Modell.

9.6 Ergebnisse - Validierung der Skala Morph

Die Skala Morph erfasst die Fähigkeit, vererbte silbenschriftliche Informationen in flektierten und abgeleiteten Formen herleiten zu können und die richtige Anwendung von Flexionsmorphemen.

9.6.1 Itemanalysen auf Basis der klassischen Testtheorie

Die Ergebnisse der Itemanalysen auf Basis der Klassischen Testtheorie zeigen (vgl. Tab. 9.13), dass die durchschnittliche Itemschwierigkeit der Skala Morph eher leicht war. Zum dritten Messzeitpunkt wurden durchschnittlich 58% der Items richtig gelöst. Zum vierten Messzeitpunkt wurden durchschnittlich 6% mehr Items richtig gelöst (vgl. Tab. 9.13). Die Ergebnisse eines t-Tests bestätigt, dass der Lernzuwachs zwischen den Messzeitpunkten signifikant ($p < .01$) ist. Die Items differenzieren gut ($r_{it3} = .42$; $r_{it4} = .43$) zwischen Personen mit höheren und Personen mit niedrigeren Merkmalsausprägungen. Die interne Konsistenz der Skala Morph ($\alpha_3 = .86$, $\alpha_4 = .87$; $r_{tt3+4} = .88$) ist hoch.

Tabelle 9.13: Itemanalysen KTT - Skala Morph

Zeitpunkt	Itemschwierigkeit		Trennschärfe		interne Konsistenz	
	$M(SD)$	Min-Max	$M(SD)$	Min-Max	α	Split-Half
MZP3	.58 (.15)	.32-.88	.42 (.13)	.19-.59	.86	.88
MZP4	.64 (.16)	.42-.88	.43 (.15)	.10-.66	.87	.88

9.6.2 Itemanalysen auf Basis der Item-Response-Theorie

9.6.2.1 Schätzung der Modellparameter

Die Ergebnisse der Skalierung der Items am eindimensionalen Rasch-Modell zum dritten und vierten Messzeitpunkt (vgl. Tab. 9.14) zeigen, dass mit einem durchschnittlichen InfitMNSQ$_3$ von .99 und OutfitMNSQ$_4$ zwischen .93 und .99 eine sehr gute Passung der Items zum Rasch-Modell bestätigt werden kann ($.75 \leq$ Infit/Outfit≤ 1.3). Die Werte liegen sehr nahe am Erwartungswert von 1 (Bond et al., 2020). Die Items differenzieren gut ($r_{it_3} = .45$, $r_{it_4} = .47$) zwischen Personen mit höheren und Personen mit niedrigeren Merkmalsausprägungen.

Der Ursprungsitempool der Skala Morph beinhaltet 24 Items. Zum vierten Messzeitpunkt wurden drei Items vor den Berechnungen ausgeschlossen, da sie keine ausreichende Passung ($.75 \leq$ Infit ≤ 1.3) zum Rasch-Modell aufweisen.

Die Outfit-Werte von sieben Items der Skala Morph liegen nicht im Bereich von $.75 \leq$ Outfit ≤ 1.3. Sie werden in den folgenden Berechnungen nicht ausgeschlossen, da der

durchschnittliche Outfit MNSQ-Wert auf eine sehr gute Passung der Items hinweist und die Ausreißer auf die Datenstruktur zurückzuführen sind und kein systematische Fehler der Skala angenommen wird. Der Ausgangsitempool zur Berechnung der Modellparameter besteht zum vierten Messzeitpunkt aus 21 Items.

Tabelle 9.14: Itemanalysen IRT - Skala Morph

Zeitpunkt	Itemschwierigkeit		Trennschärfe		InfitMNSQ		OutfitMNSQ	
	$M(SD)$	Min-Max	$M(SD)$	Min-Max	$M(SD)$	Min-Max	$M(SD)$	Min-Max
MZP3	.01 (.91)	-2.01-1.53	.47 (.13)	.21-.64	.99 (.15)	.8-1.3	.99 (.26)	.69-1.63
MZP4	0 (1.01)	-1.73-1.37	.45 (.17)	.14-.69	.99 (.20)	.69-1.37	.93 (.34)	.50-1.9

9.6.2.2 Vergleich der Item- und Personenparameter

Die Items der Skala Morph erfassen die unterschiedlichen Personenfähigkeiten der Schüler*innen gut (EAP Rel.$_3$: .81; WLE Rel.$_3$: .76; EAP Rel.$_4$: .80; WLE Rel.$_4$: .74). Der Vergleich der Item- und Personenparameter zeigt (vgl. Abb. 9.13; 9.14), dass die Items über den gesamten Bereich der Personenfähigkeiten streuen. Zu beiden Messzeitpunkten werden die Personenfähigkeiten der Schüler*innen durch die Items an den Randbereichen des Fähigkeitsniveaus nicht ausreichend abgedeckt.

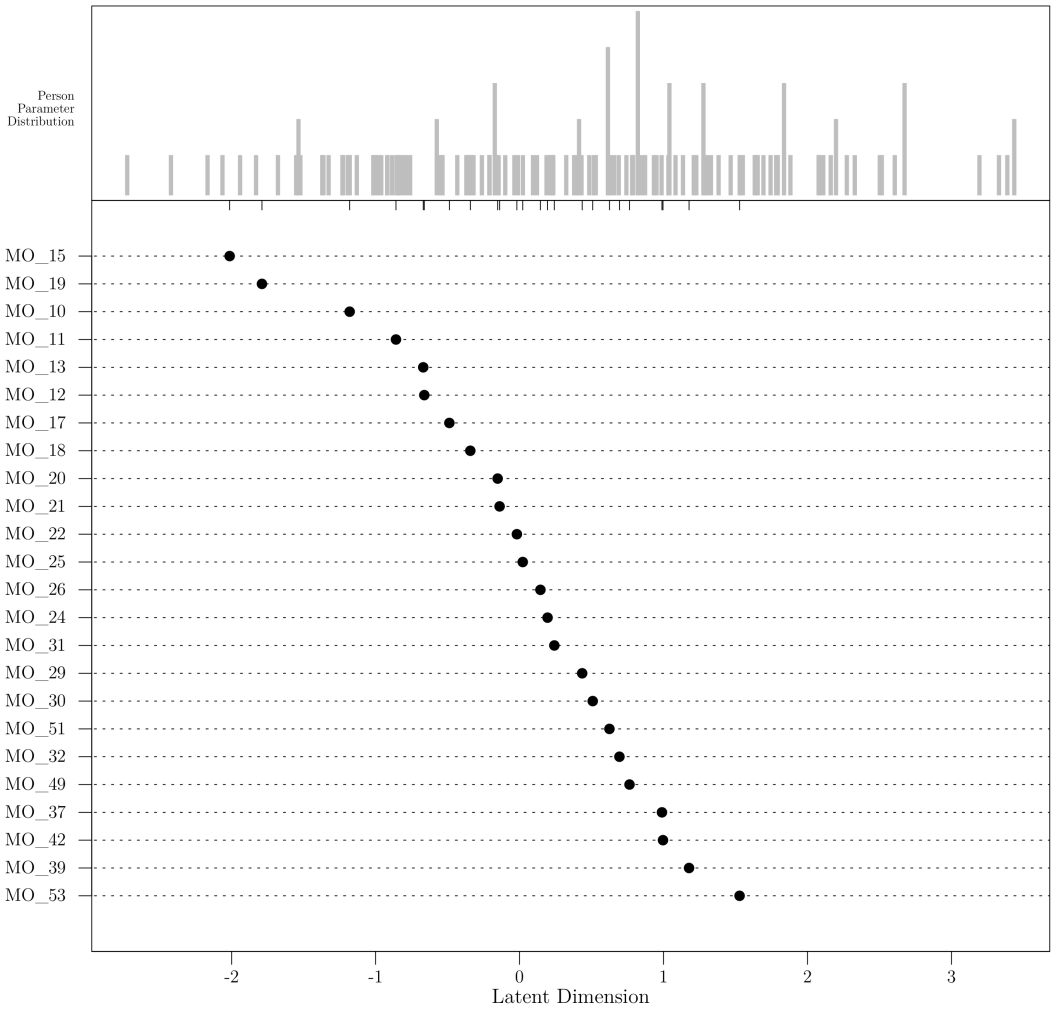

Abbildung 9.13: Person-Item Map. Verteilung der Personenparameter und Position der Items auf latenter Dimension. Skala Morph; MZP3.

Die Verteilung der Itemschwierigkeiten auf der latenten Dimension im unteren Bereich der Person-Item-Map (vgl. Abb. 9.14) zeigt, dass die Items über den mittleren Bereich der Personenfähigkeiten streuen. Die Verteilung der Personenfähigkeit ist asymmetrisch, der Gipfel liegt auf der rechten Seite, viele Personen haben einen hohen Personenscore erreicht. Es fehlen Items für leistungsschwache und leistungsstarke Schüler*innen. Die

übergeordnete Messgenauigkeit der Skala Morph ist gut ($EAPRel.3 : .81; WLERel.3 : .76$).

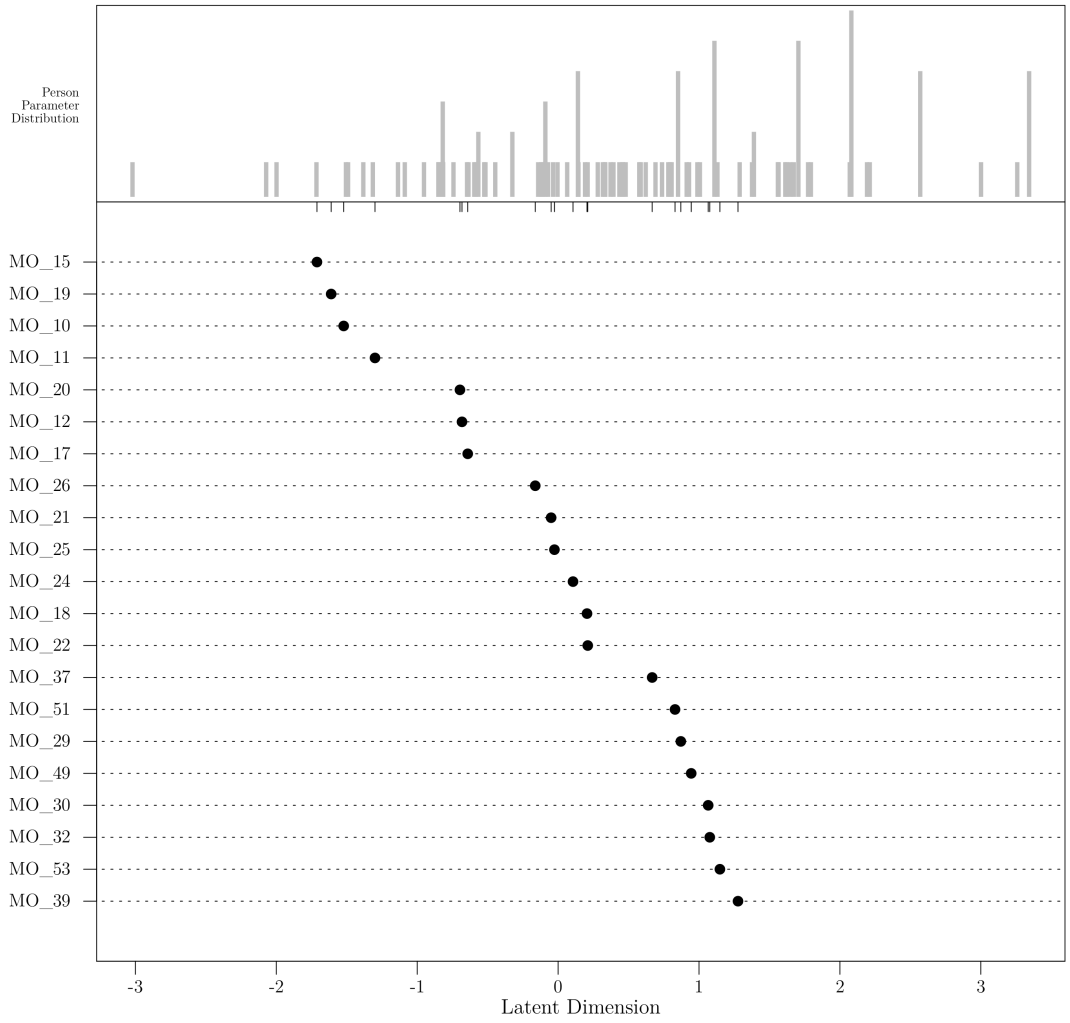

Abbildung 9.14: Person-Item Map. Verteilung der Personenparameter und Position der Items auf latenter Dimension. Skala Morph; MZP4. LA = 24.

9.6.2.3 Prüfung der Itemhomogenität

Die Ergebnisse des Likelihood-Quotienten-Tests der Skala Morph [9] zeigen, dass sich die Schätzungen der Item-Parameter nicht signifikant aufgeteilt nach dem Median unterscheiden und keine Modellverletzungen ($p < .01$) vorliegen (vgl. Tab. 9.15, 9.15, 9.16).

[9] Einzelne Items weisen keine passende Antwortmuster zur Durchführung des Likelihood-Quotienten-Tests auf und werden automatisch von den Analysen ausgeschlossen. Der Ausgangsitempool zur Prüfung der Itemhomogenität besteht zum dritten Messzeitpunkt und zum vierten Messzeitpunkt aus 20 Items.

Tabelle 9.15: Itemhomogenität - Skala Morph

	LU	LA	LR	*df*	*p*
MZP3	24	20	31.23	19	.04
MZP4	24	20	31.20	19	.04

Anmerkungen. Ergebnisse des Likelihood-Quotienten-Tests für Skala Morph.; Teilungskriterium Median; Bonferroni-Korrektur ($p - Wert < .01$).
LU = Anzahl der Lupenstellen im Ursprungsitempool; LA = Anzahl der Lupenstellen, die auf Itemhomogenität geprüft werden konnten.

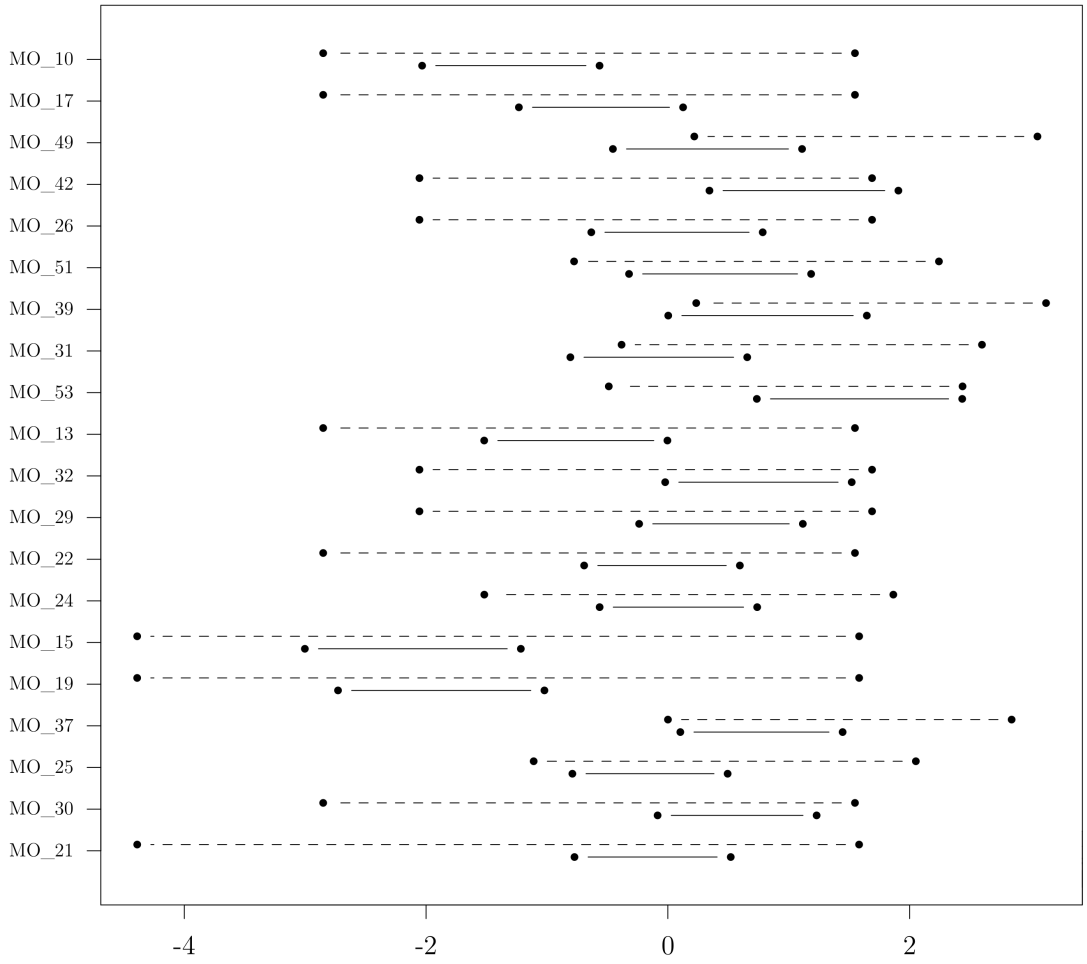

Abbildung 9.15: Plot der Konfidenzintervalle. Itemschwierigkeiten aufgeteilt nach Teilungskriterium Median; Bonferroni-Korrektur. Skala Morph; MZP3.

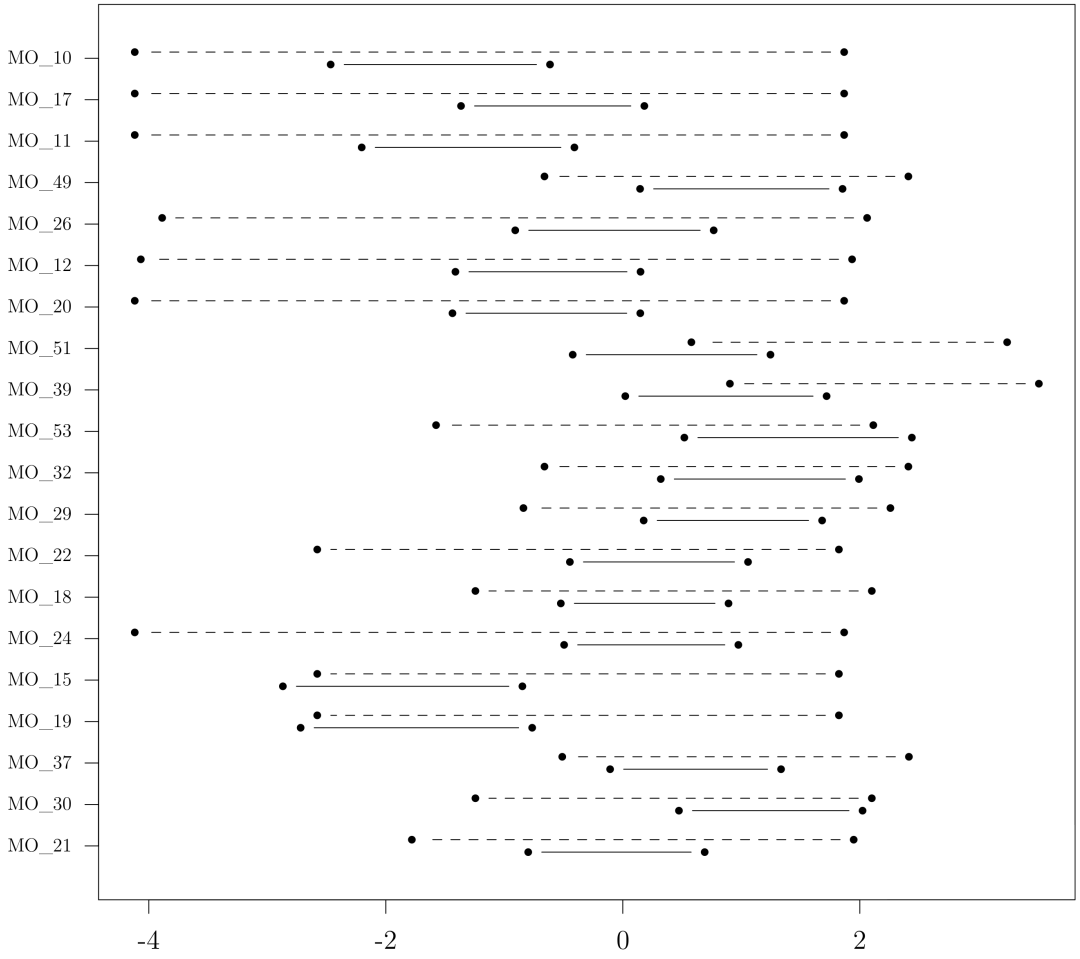

Abbildung 9.16: Plot der Konfidenzintervalle. Itemschwierigkeiten aufgeteilt nach Teilungskriterium Median; Bonferroni-Korrektur. Skala Morph; MZP4.

9.6.2.4 Prüfung der Testfairness

Die Ergebnisse des Likelihood-Quotienten-Tests zur Prüfung der Testfairness der Skala Morph [10] (vgl. Tab. 9.16) zeigen, dass sich die Schätzungen der Item-Parameter nicht signifikant aufgeteilt nach dem Geschlecht unterscheiden und kein ($p < .01$) Differential Item Functioning (DIF) vorliegt. Schüler*innen mit gleichem Fähigkeitsniveau erreichen die gleichen Personenscores im ReKoMe (vgl. Tab. 9.17, 9.18).

[10] Einzelne Items weisen keine passende Antwortmuster zur Durchführung des Likelihood-Quotienten-Tests auf und werden automatisch von den Analysen ausgeschlossen. Der Ausgangsitempool zur Prüfung der Testfairness besteht zum dritten Messzeitpunkt aus 24 Items und zum vierten Messzeitpunkt aus 21 Items.

Tabelle 9.16: Testfairness - Skala Morph

	LU	LA	LR	*df*	*p*
MZP3	24	24	16.61	23	.83
MZP4	24	21	13.72	20	.84

Anmerkungen. Ergebnisse des Likelihood-Quotienten-Tests für Skala Morph.; Teilungskriterium Geschlecht; Bonferroni-Korrektur ($p - Wert < .01$).
LU = Anzahl der Lupenstellen im Ursprungsitempool; LA = Anzahl der Lupenstellen, die auf Itemhomogenität geprüft werden konnten.

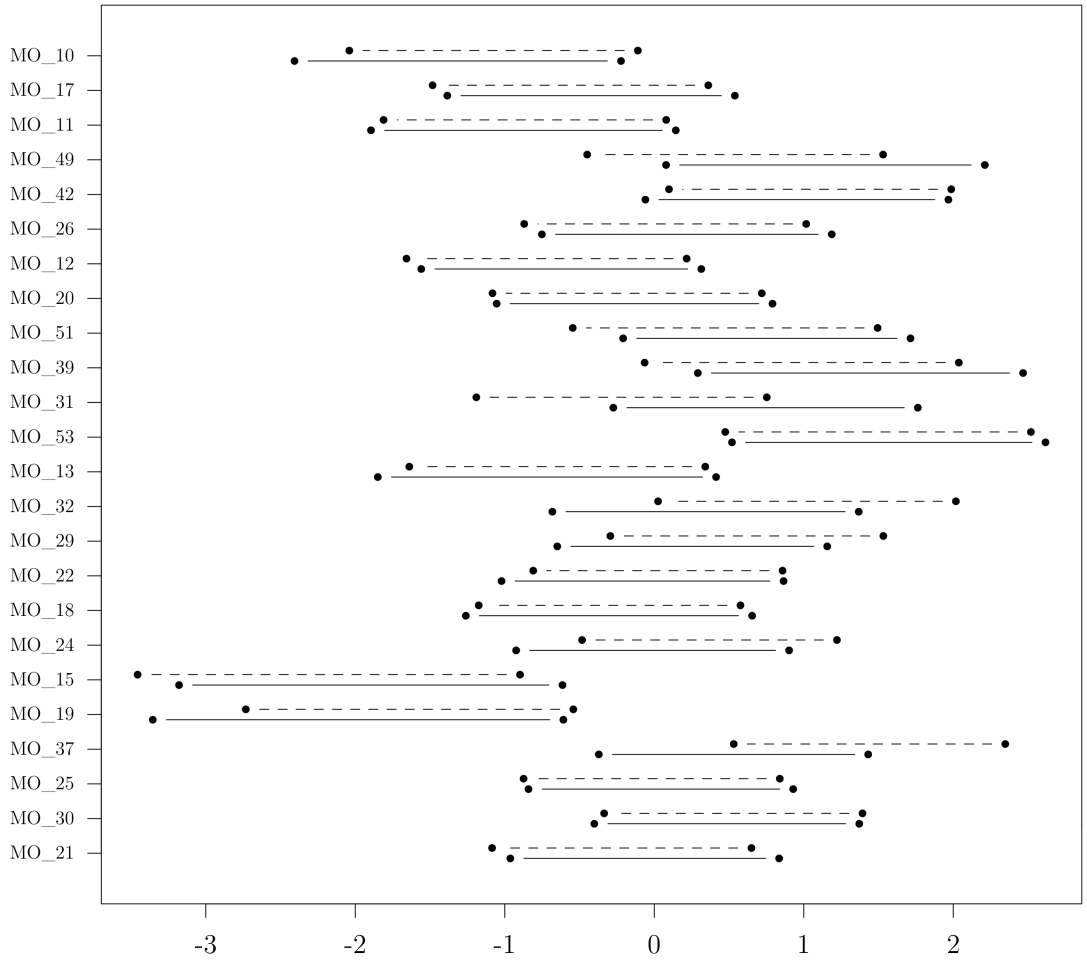

Abbildung 9.17: DIF-Plot. Itemschwierigkeiten aufgeteilt nach Geschlecht mit korrigierten Konfidenzintervall; Bonferroni-Korrektur. Skala Morph; MZP 3.

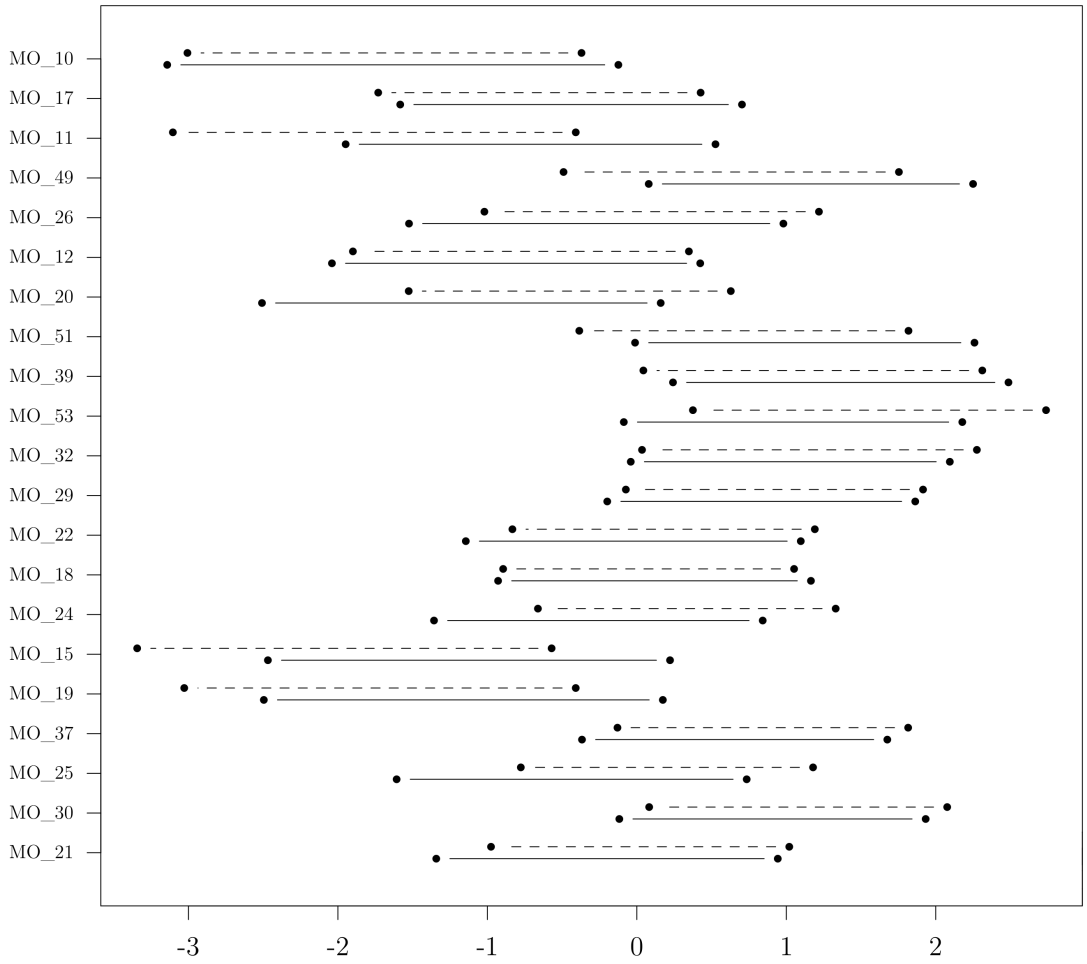

Abbildung 9.18: DIF-Plot. Itemschwierigkeiten aufgeteilt nach Geschlecht mit korrigierten Konfidenzintervall; Bonferroni-Korrektur. Skala Morph; MZP 4.

9.6.2.5 Modellvergleich

Der Modellvergleich der Skala Morph zwischen dem Rasch-Modell und dem Birnbaum-Modell bestätigt, dass die Daten besser durch das Rasch-Modell ($p < .001$) erklärt werden. Der BIC-Wert zeigt (vgl. Tab. 9.17) jeweils eine bessere Datenbeschreibung mit einem niedrigeren Wert an.

Tabelle 9.17: Modellvergleich - Skala Morph

Modell	MZP3		MZP4	
	AIC	BIC	AIC	BIC
1 PL	3179.43	3253.33	2462	2526.49
2 PL	3160.91	3302.79	2567.24	2701.83

Anmerkungen. 1 PL = Rasch-Modell; 2 PL = Birnbaum-Modell.

9.7 Ergebnisse - Validierung der Skala Peri

Die Skala Peri erfasst die Fähigkeit, Markierungen in offenen Silben setzen und vererbte Schreibweisen herleiten zu können und Lernwörter und Fremdwortschreibungen richtig zu schreiben.

9.7.1 Itemanalysen auf Basis der klassischen Testtheorie

Die Ergebnisse der Itemanalysen auf Basis der Klassischen Testtheorie für die Skala Peri [11] zeigen (vgl. Tab. 9.18), dass die durchschnittliche Itemschwierigkeit zum dritten Messzeitpunkt im mittleren Schwierigkeitsbereich liegt, es wurden durchschnittlich 50% der Items richtig gelöst. Zum vierten Messzeitpunkt wurden durchschnittlich 4% mehr Items richtig gelöst. Der Lernzuwachs ist nicht signifikant ($p > .05$). Die Items differenzieren gut ($r_{it3} = .36$; $r_{it4} = .44$) zwischen Personen mit höheren und Personen mit niedrigeren Merkmalsausprägungen. Die interne Konsistenz der Skala Morph ($\alpha_3 = .73$, $\alpha_4 = .81$; $r_{tt3} = .81$, $r_{tt4} = .92$) liegt im akzeptablen bis guten Bereich.

Tabelle 9.18: Itemanalysen KTT - Skala Peri

Zeitpunkt	Itemschwierigkeit		Trennschärfe		interne Konsistenz	
	$M(SD)$	Min-Max	$M(SD)$	Min-Max	α	Split-Half
MZP3	.50 (.29)	.10-.92	.36 (.11)	.19-.51	.73	.81
MZP4	.54 (.29)	.11-.96	.44 (.20)	.06-.72	.81	.92

9.7.2 Itemanalysen auf Bais der Item-Response-Theorie

9.7.2.1 Schätzung der Modellparameter

Die Ergebnisse der Skalierung der Items am eindimensionalen Rasch-Modell zum dritten und vierten Messzeitpunkt (vgl. Tab. 9.19) zeigen, dass mit einem durchschnittlichen InfitMNSQ von .85-.89 und OutfitMNSQ zwischen .94 und 1.11 eine sehr gute Passung der Items zum Rasch-Modell bestätigt werden kann ($.75 \leq Infit/Outfit \leq 1.3$). Die Werte liegen sehr nahe am Erwartungswert von 1 (Bond et al., 2020). Die Items differenzieren gut ($r_{it_3} = .40$, $r_{it_4} = .46$) zwischen Personen mit höheren und Personen mit niedrigeren Merkmalsausprägungen.

Der Ursprungsitempool der Skala Peri beinhaltet 13 Items. Vor den Berechnungen wurde zum dritten Messzeitpunkt ein Item und zum vierten Messzeitpunkt 4 Items ausgeschlossen, da die Items nicht raschkonform (InfitMNSQ $< .75$) sind. Im Itempool befinden

[11] In der Skala sind insgesamt 13 Items enthalten. Für die Durchführbarkeit der Analysen musste ein Item von den Berechnungen ausgeschlossen werden.

sich zu beiden Messzeitpunkten auch Items, dessen Outfit-Werte (z.B. MZP3: Min-Max = .32-1.89) nicht im Bereich von ($.75 \leq$ Outfit ≤ 1.3) liegen. Die gemittelten Werte des OutfitMNSQ-Werts liegen zu beiden Zeitpunkten (OutfitMNSQ_3: .94, OutfitMNSQ_4: 1.11) sehr nahe am Erwartungswert von 1 und werden nicht von den Berechnungen ausgeschlossen, da von keinem systematischen Fehler der Skala ausgegangen wird. Der Ausgangsitempool zur Berechnung der Modellparameter besteht zum dritten Messzeitpunkt aus 12 Items und zum vierten Messzeitpunkt aus 9 Items.

Tabelle 9.19: Itemanalysen IRT- Skala Peri

Zeitpunkt	Itemschwierigkeit		Trennschärfe		InfitMNSQ		OutfitMNSQ	
	$M(SD)$	Min-Max	$M(SD)$	Min-Max	$M(SD)$	Min-Max	$M(SD)$	Min-Max
MZP3	-.10 (2.0)	-3.02-3.01	.40 (.11)	.22-.57	.89 (.13)	.67-1.1	.94 (.5)	.32-1.89
MZP4	-.12 (2.3)	-3.83-3.66	.46 (.19)	.09-.74	.85 (.17)	.55-1.16	1.11 (.86)	.56-3.27

9.7.2.2 Vergleich der Item- und Personenparameter

Die Items der Skala Peri erfassen die unterschiedlichen Personenfähigkeiten der Schüler*innen je nach Gütemaß akzeptabel bis fragwürdig (EAP Rel.$_3$: .67; WLE Rel.$_3$: .61; EAP Rel.$_4$: .71; WLE Rel.$_4$: .59). Der Vergleich der Item- und Personenparameter zeigt (vgl. Abb. 9.7; 9.8), dass die Items über den mittleren Bereich der Personenfähigkeiten streuen. Zu beiden Messzeitpunkten werden die Personenfähigkeiten der Schüler*innen durch die Items an den Randbereichen des Fähigkeitsniveaus nicht ausreichend abgedeckt.

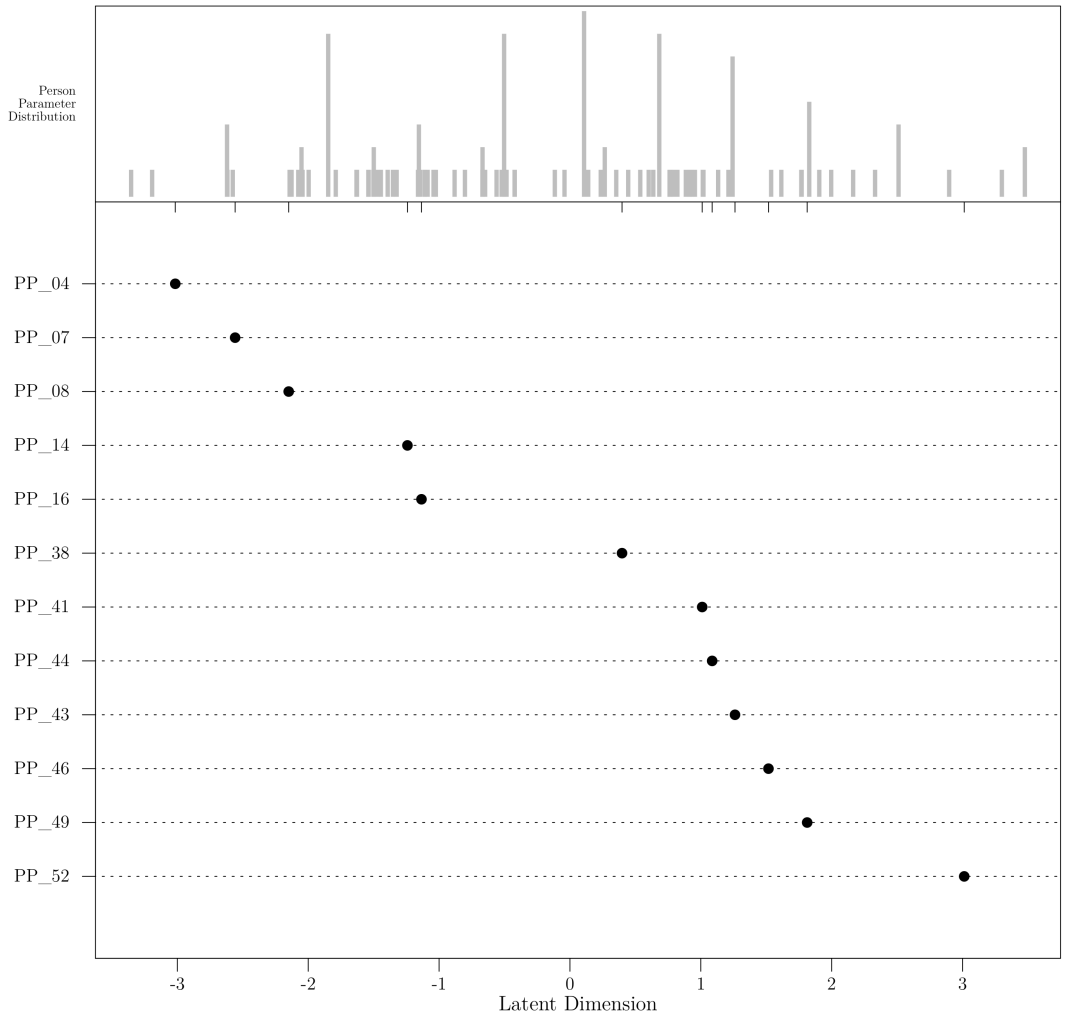

Abbildung 9.19: Person-Item Map. Verteilung der Personenparameter und Position der Items auf latenter Dimension. Skala Peri; MZP3. LA= 12.

Die Verteilung der Itemschwierigkeiten auf der latenten Dimension im unteren Bereich der Person-Item Map zeigt, dass die Items über den mittleren Bereich der Personenfähigkeiten streuen. Es fehlen Items für leistungsschwache und leistungsstarke Schüler*innen. Die übergeordnete Messgenauigkeit der Skala Peri ist je nach Gütemaß akzeptabel bis fragwürdig (EAP Rel.$_3$: .67; WLE Rel.$_3$: .61).

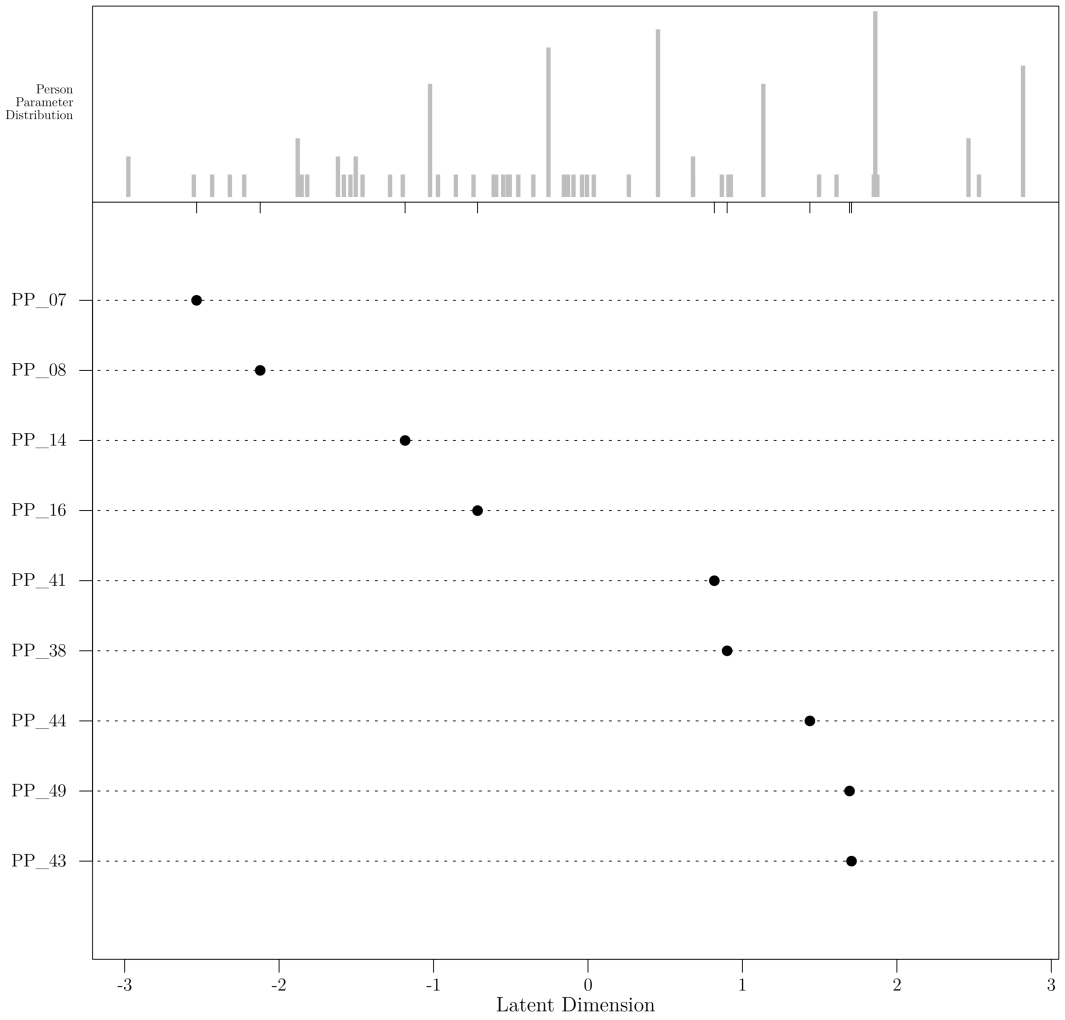

Abbildung 9.20: Person-Item Map. Verteilung der Personenparameter und Position der Items auf latenter Dimension. Skala Peri; MZP4. LA= 9.

Zum vierten Messzeitpunkt haben sehr viele Personen einen hohen Personenscore erreicht. Es fehlen ebenfalls Items für leistungsschwache und leistungsstarke Schüler*innen. Die übergeordnete Messgenauigkeit ist der Skala Peri ist je nach Gütemaß akzeptabel bis fragwürdig (EAP Rel.$_4$: .71; WLE Rel.$_4$: .59). Zu beiden Messzeitpunkten fehlen Items im mittleren Schwierigkeitsbereich.

9.7.2.3 Prüfung der Itemhomogenität

Die Ergebnisse des Likelihood-Quotienten-Tests der Skala Peri [12] zeigen, dass sich die Schätzungen der Item-Parameter nicht signifikant aufgeteilt nach dem Median unter-

[12] Einzelne Items weisen keine passende Antwortmuster zur Durchführung des Likelihood-Quotienten-Tests auf und werden automatisch von den Analysen ausgeschlossen. Der Ausgangsitempool zur Prüfung der Itemhomogenität besteht zum dritten Messzeitpunkt und zum vierten Messzeitpunkt aus 9 Items.

scheiden und keine Modellverletzungen ($p < .01$) vorliegen (vgl. Tab. 9.20, Abb. 9.21, 9.22).

Tabelle 9.20: Itemhomogenität - Skala Peri

	LU	LA	LR	*df*	*p*
MZP3	13	9	6.585	8	.58
MZP4	13	9	5.393	8	.72

Anmerkungen. Ergebnisse des Likelihood-Quotienten-Tests für Skala Peri.; Teilungskriterium Median; Bonferroni-Korrektur ($p - Wert < .01$).
LU = Anzahl der Lupenstellen im Ursprungsitempool; LA = Anzahl der Lupenstellen, die auf Itemhomogenität geprüft werden konnten.

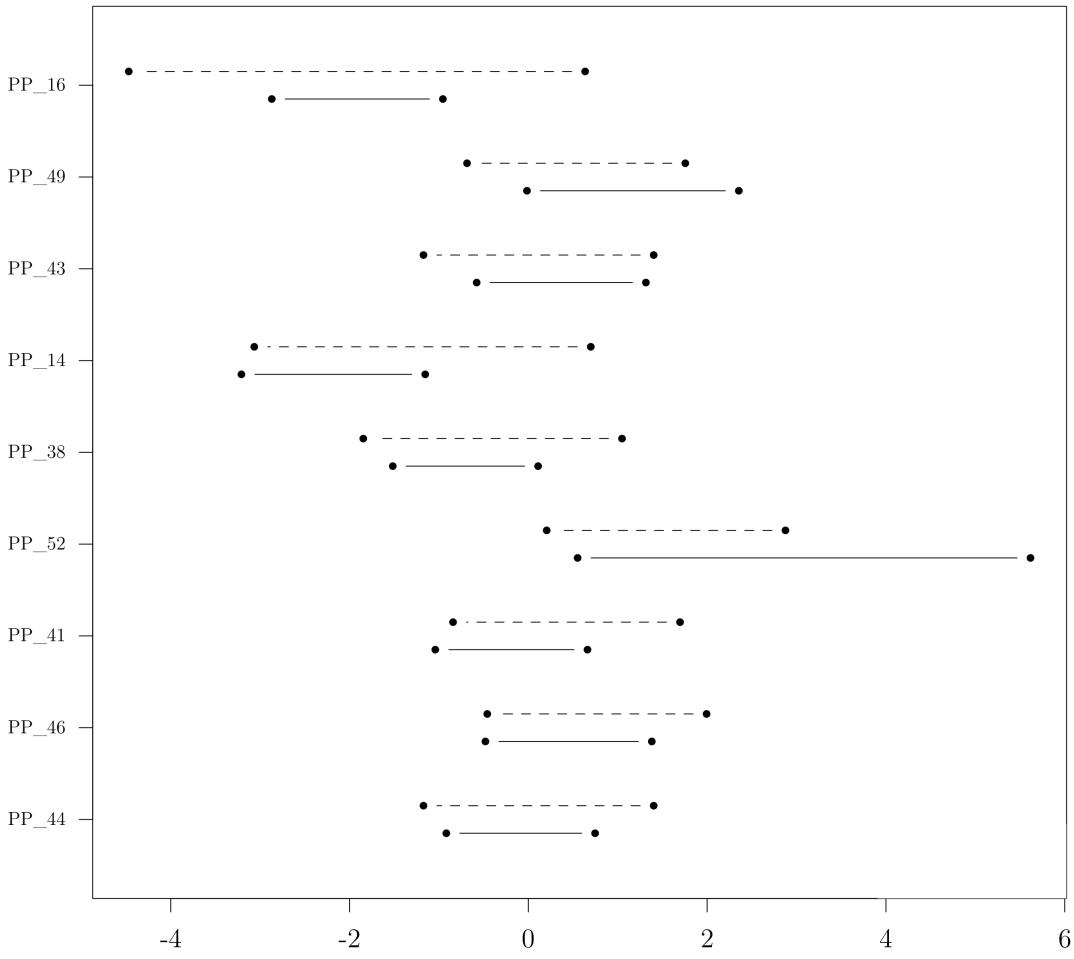

Abbildung 9.21: Plot der Konfidenzintervalle. Itemschwierigkeiten aufgeteilt nach Teilungskriterium Median; Bonferroni-Korrektur. Skala Peri; MZP3.

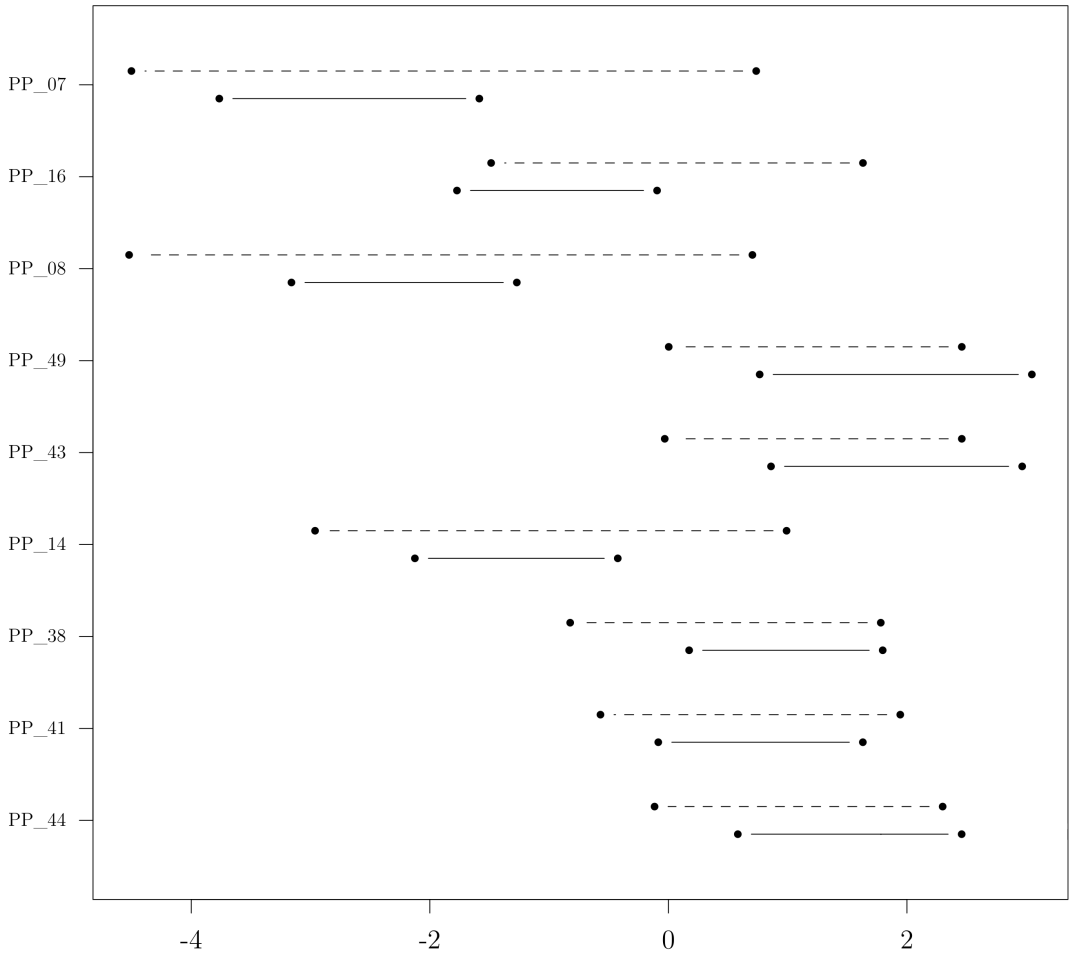

Abbildung 9.22: Plot der Konfidenzintervalle. Itemschwierigkeiten aufgeteilt nach Teilungskriterium Median; Bonferroni-Korrektur. Skala Peri; MZP4.

9.7.2.4 Prüfung der Testfairness

Die Ergebnisse des Likelihood-Quotienten-Tests zur Prüfung der Testfairness (vgl. Tab. 9.16) der Skala Peri [13] zeigen, dass sich die Schätzungen der Item-Parameter nicht signifikant aufgeteilt nach dem Geschlecht unterscheiden und kein Differential Item Functioning (DIF) ($p < .01$) vorliegt. Schüler*innen mit gleichem Fähigkeitsniveau erreichen die gleichen Personenscores im ReKoMe (vgl. Abb. 9.23, 9.24).

[13] Einzelne Items weisen keine passende Antwortmuster zur Durchführung des Likelihood-Quotienten-Tests auf und werden automatisch von den Analysen ausgeschlossen. Der Ausgangsitempool zur Prüfung der Testfairness besteht zum dritten Messzeitpunkt aus 12 Items und zum vierten Messzeitpunkt aus 9 Items.

Tabelle 9.21: Testfairness - Skala Peri

	LU	LA	LR	*df*	*p*
MZP3	12	12	16.95	11	.11
MZP4	12	9	12.25	8	.14

Anmerkungen. Ergebnisse des Likelihood-Quotienten-Tests für Skala Peri.; Teilungskriterium Geschlecht; Bonferroni-Korrektur ($p - Wert < .01$).
LU = Anzahl der Lupenstellen im Ursprungsitempool; LA = Anzahl der Lupenstellen, die auf Itemhomogenität geprüft werden konnten.

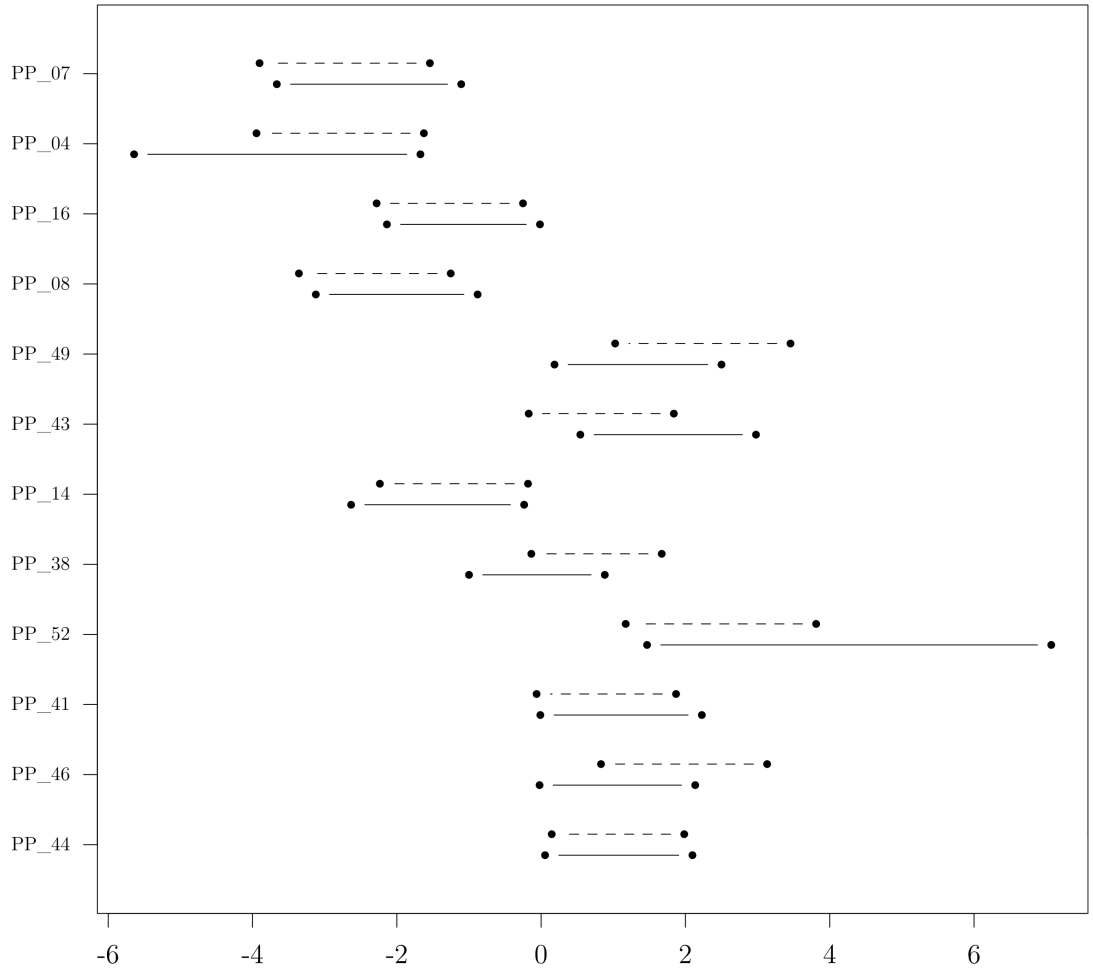

Abbildung 9.23: DIF-Plot. Itemschwierigkeiten aufgeteilt nach Geschlecht mit korrigierten Konfidenzintervall; Bonferroni-Korrektur. Skala Peri; MZP 3.

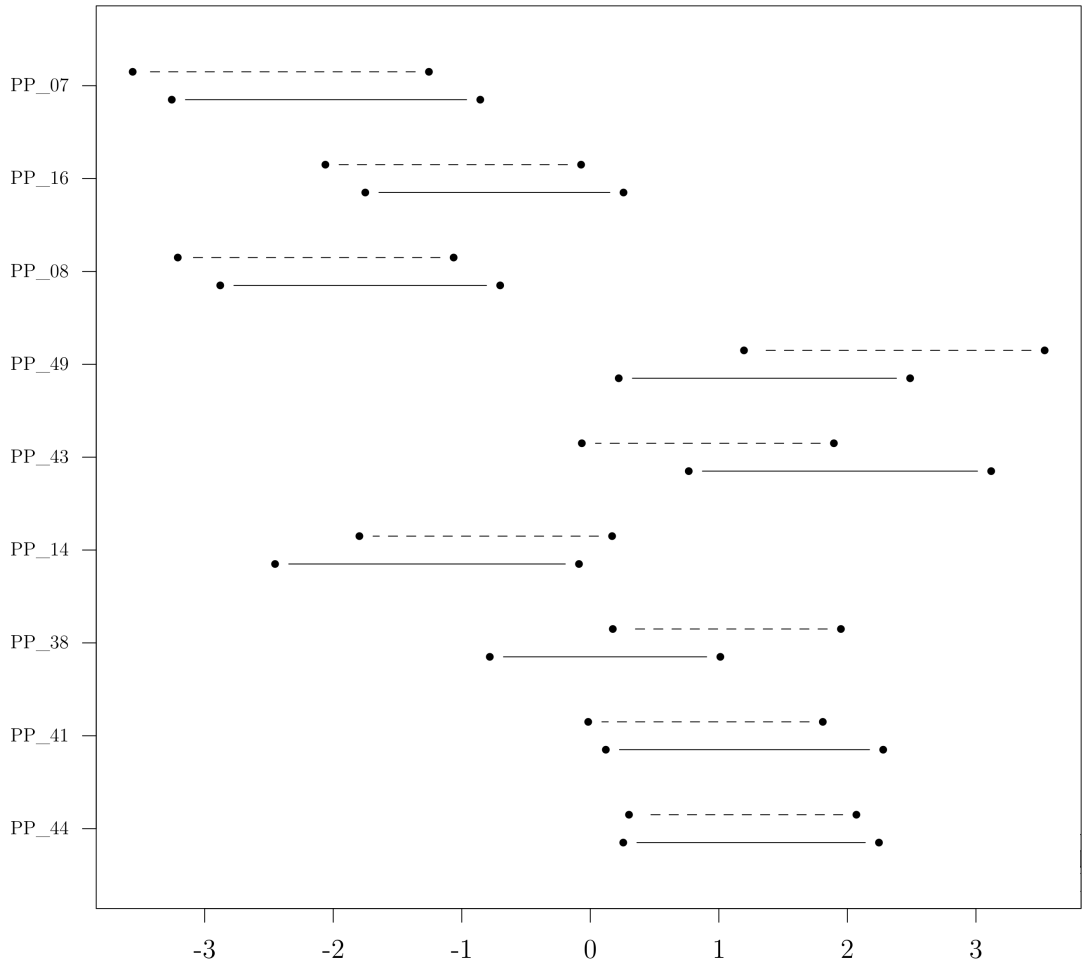

Abbildung 9.24: DIF-Plot. Itemschwierigkeiten aufgeteilt nach Geschlecht mit korrigierten Konfidenzintervall; Bonferroni-Korrektur. Skala Peri; MZP 4.

9.7.2.5 Modellvergleich

Der Modellvergleich der Skala Peri zwischen dem Rasch-Modell und dem Birnbaum-Modell bestätigt, dass die Daten besser durch das Rasch-Modell ($p < .001$) erklärt werden. Der BIC-Wert zeigt (vgl. Tab. 9.17) jeweils eine bessere Datenbeschreibung mit einem niedrigeren Wert an.

Tabelle 9.22: Modellvergleich - Skala Peri

Modell	MZP3		MZP4	
	AIC	BIC	AIC	BIC
1 PL	1247.68	1288.04	1034.66	1073.09
2 PL	1249.62	1324.58	1027.13	1098.50

Anmerkungen. 1 PL = Rasch-Modell; 2 PL = Birnbaum-Modell.

9.8 Ergebnisse - Validierung der Skala Wortbil

Die Skala Wortbil erfasst die Fähigkeit, Wortarten und Wortbildungsmorpheme zu kennen und in Ableitungen und Komposita richtig anzuwenden.

9.8.1 Itemanalysen auf Basis der klassischen Testtheorie

Die Ergebnisse der Itemanalysen [14] auf Basis der Klassischen Testtheorie zeigen (vgl. Tab. 9.23), dass die durchschnittliche Itemschwierigkeit zu beiden Messzeitpunkten schwieriger waren, es wurden 44% der Items richtig gelöst. Die Items trennen mit einem durchschnittlichen Trennschärfekoeffizienten von $r_{it3} = .22$ bis $r_{it4} = .32$ nicht im ausreichenden Maße zwischen Personen mit höheren und Personen mit niedrigeren Merkmalsausprägungen. Die interne Konsistenz der Skala Wortbil ist nicht ausreichend ($\alpha_{3+4} = .38$; $r_{tt3+4} = .27$). Die Werte müssen jedoch auch im Zusammenhang mit der Anzahl der geringen Anzahl von Items (Lup = 4) interpretiert werden (Schermelleh-Engel & Gäde, 2020).Der Wert Cronbachs Alpha steht im Zusammenhang mit der Anzahl der Items. Je mehr Items im Test enthalten sind, desto höher wird der Wert der Reliabilität der Skala (Schermelleh-Engel & Gäde, 2020).

Tabelle 9.23: Itemanalysen KTT - Skala Wortbil

Zeitpunkt	Itemschwierigkeit		Trennschärfe		interne Konsistenz	
	$M(SD)$	Min-Max	$M(SD)$	Min-Max	α	Split-Half
MZP3	.44 (.20)	.21-.68	.22 (.18)	.09-.48	.38	.27
MZP4	.44 (.19)	.23-.64	.32 (.17)	.09-.49	.38	.27

9.8.2 Itemanalysen auf Basis der Item-Response-Theorie

9.8.2.1 Schätzung der Modellparameter

Die Ergebnisse der Skalierung der Items am eindimensionalen Rasch-Modell zum dritten und vierten Messzeitpunkt (vgl. Tab. 9.24) zeigen, dass mit einem durchschnittlichen InfitMNSQ und OutfitMNSQ von .98 bis .99 eine sehr gute Passung der Items zum Rasch-Modell bestätigt werden kann ($.75 \leq$ Infit/Outfit≤ 1.3). Die Werte liegen sehr nahe am Erwartungswert von 1 (Bond et al., 2020). Die Trennschärfe konnte aus statistischen Gründen nicht berechnet werden, da der Datensatz dieser Skala zu viele fehlende Daten enthält. Der Ursprungsitempool der Skala Wortbil beinhaltet 5 Items. Vor den Berechnungen wurde zum dritten Messzeitpunkt ein Item ausgeschlossen, da es nicht raschkonform

[14] In der Skala sind insgesamt 5 Items enthalten. Für die Durchführbarkeit der Analysen musste ein Item von den Berechnungen ausgeschlossen werden.

ist (InfitMNSQ < .75). Der Ausgangsitempool zur Berechnung der Modellparameter besteht zum dritten Messzeitpunkt aus 4 Items.

Tabelle 9.24: Itemanalysen IRT- Skala Wortbil

Zeitpunkt	Itemschwierigkeit		InfitMNSQ		OutfitMNSQ	
	$M(SD)$	Min-Max	$M(SD)$	Min-Max	$M(SD)$	Min-Max
MZP3	-.01 (.36)	-.4-.29	.99 (.12)	.85-1.15	.99 (.14)	.83-1.17
MZP4	-.14 (.32)	-.46-.24	.98 (.14)	.81-1.12	.98 (.2)	.76-1.21

9.8.2.2 Vergleich der Item- und Personenparameter

Die Items der Skala Wortbil erfassen die unterschiedlichen Personenfähigkeiten der Schüler*innen nicht ausreichend (EAP Rel.$_3$: .0; WLE Rel.$_3$: .-1.81; EAP Rel.$_4$: .17; WLE Rel.$_4$: -.41). Der Vergleich der Item- und Personenparameter zeigt (vgl. Abb. 9.25; 9.26), dass die Items nur im mittleren Bereich der Personenfähigkeiten streuen. Zu beiden Messzeitpunkten werden die Personenfähigkeiten der Schüler*innen durch die Items an den Randbereichen des Fähigkeitsniveaus nicht ausreichend abgedeckt.

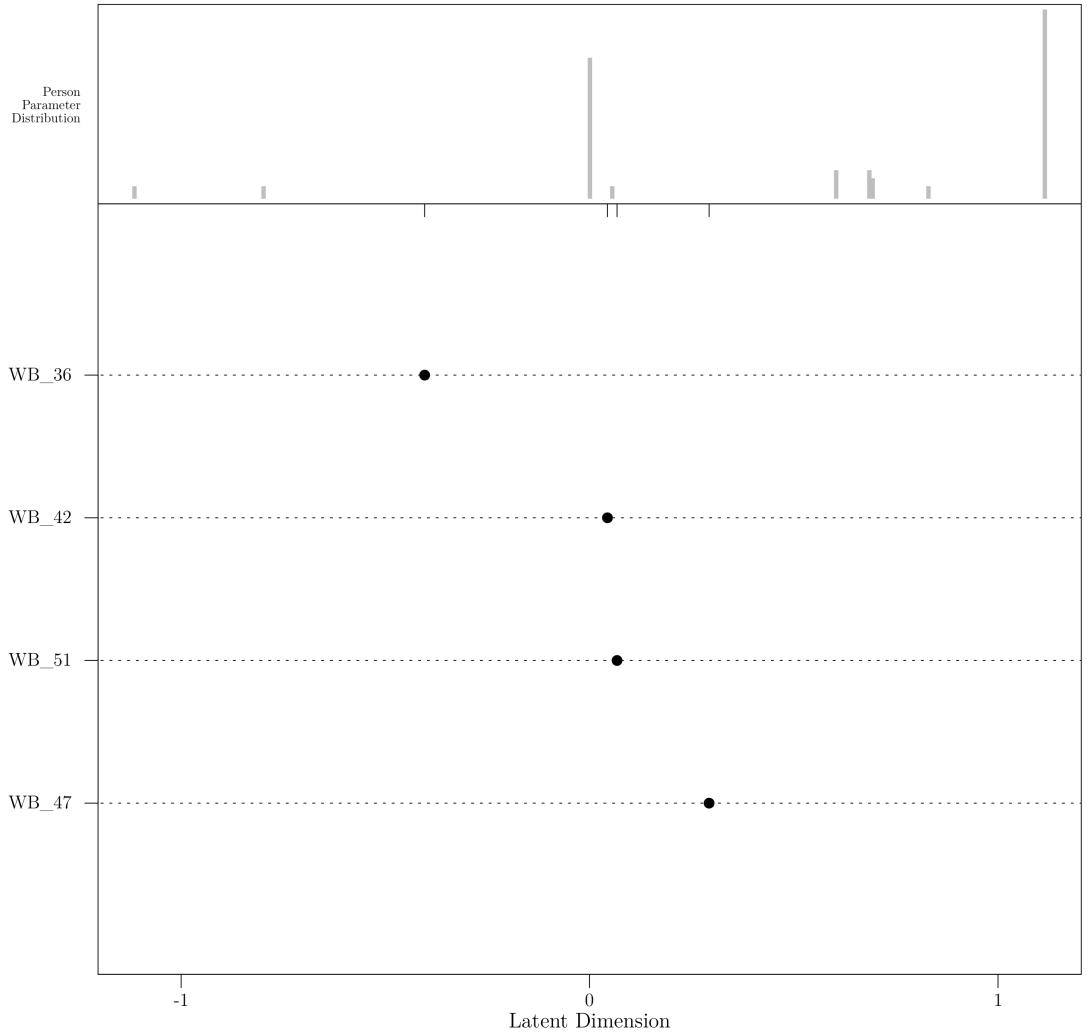

Abbildung 9.25: Person-Item Map. Verteilung der Personenparameter und Position der Items auf latenter Dimension. Skala Wortbil; MZP3. LA= 4.

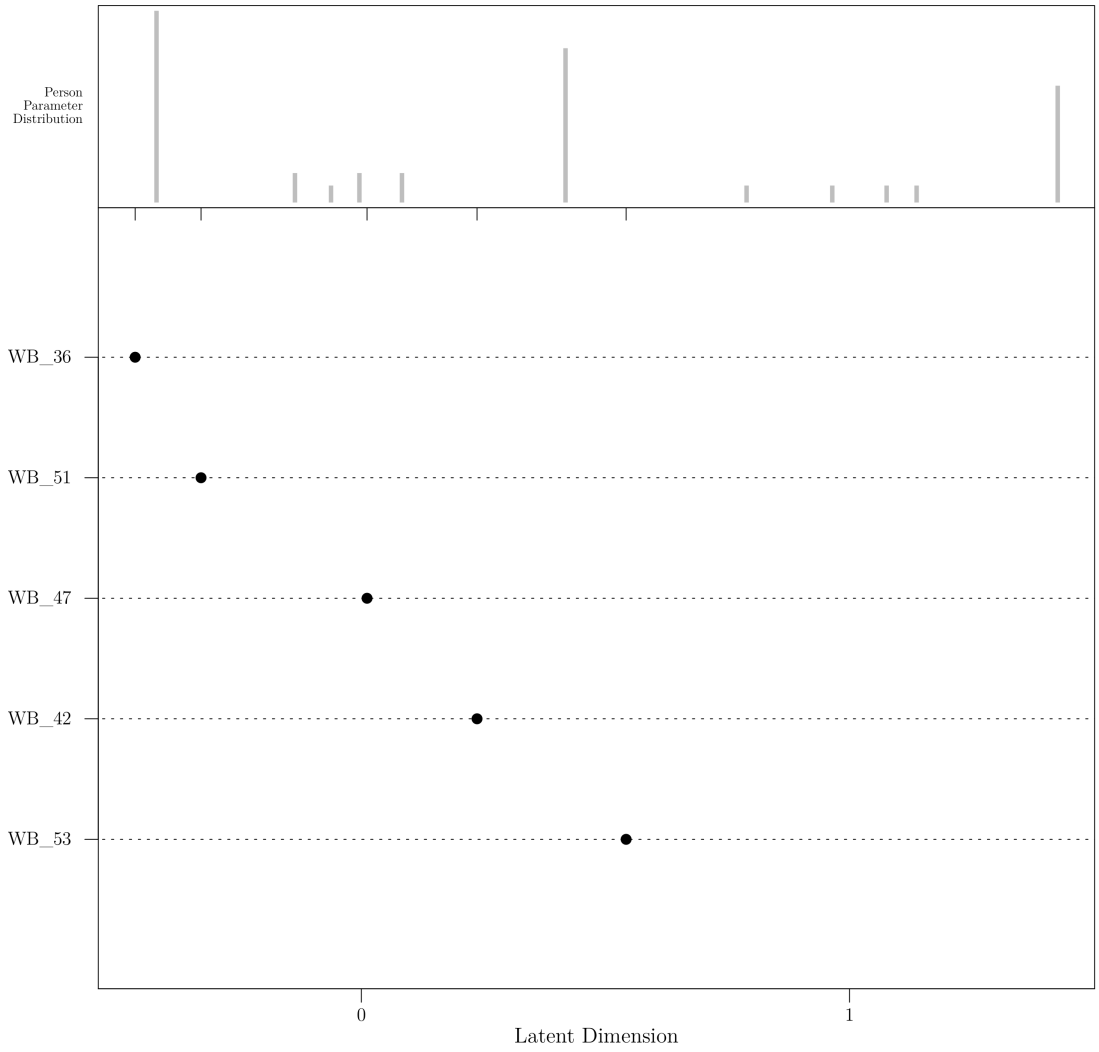

Abbildung 9.26: Person-Item Map. Verteilung der Personenparameter und Position der Items auf latenter Dimension. Skala Wortbil; MZP4.

9.8.2.3 Prüfung der Itemhomogenität

Die Durchführung eines Likelihood-Quotienten-Test zur Überprüfung der Itemhomogenität ist anhand des Teilungskriteriums Median aufgrund der Datenstruktur nicht möglich.

9.8.2.4 Prüfung der Testfairness

Die Ergebnisse des Likelihood-Quotienten-Tests zur Prüfung der Testfairness (vgl. Tab. 9.25) der Skala Wortbil [15] zeigen, dass sich die Schätzungen der Item-Parameter nicht signifikant aufgeteilt nach dem Geschlecht unterscheiden und kein Differential Item Functioning (DIF) ($p < .01$) vorliegt. Schüler*innen mit gleichem Fähigkeitsniveau erreichen die gleichen Personenscores im ReKoMe (vgl. Abb. 9.27, 9.28).

[15] Einzelne Items weisen keine passende Antwortmuster zur Durchführung des Likelihood-Quotienten-Tests auf und werden automatisch von den Analysen ausgeschlossen. Der Ausgangsitempool zur Prüfung der Testfairness besteht zum dritten Messzeitpunkt aus 4 Items und zum vierten Messzeitpunkt aus 5 Items.

Tabelle 9.25: Testfairness - Skala Wortbil

	LU	LA	LR	*df*	*p*
MZP3	5	4	2.026	3	.57
MZP4	5	5	4.766	4	.31

Anmerkungen. Ergebnisse des Likelihood-Quotienten-Tests für Skala Wortbil.; Teilungskriterium Geschlecht; Bonferroni-Korrektur ($p - Wert < .01$).
LU = Anzahl der Lupenstellen im Ursprungsitempool; LA = Anzahl der Lupenstellen, die auf Itemhomogenität geprüft werden konnten.

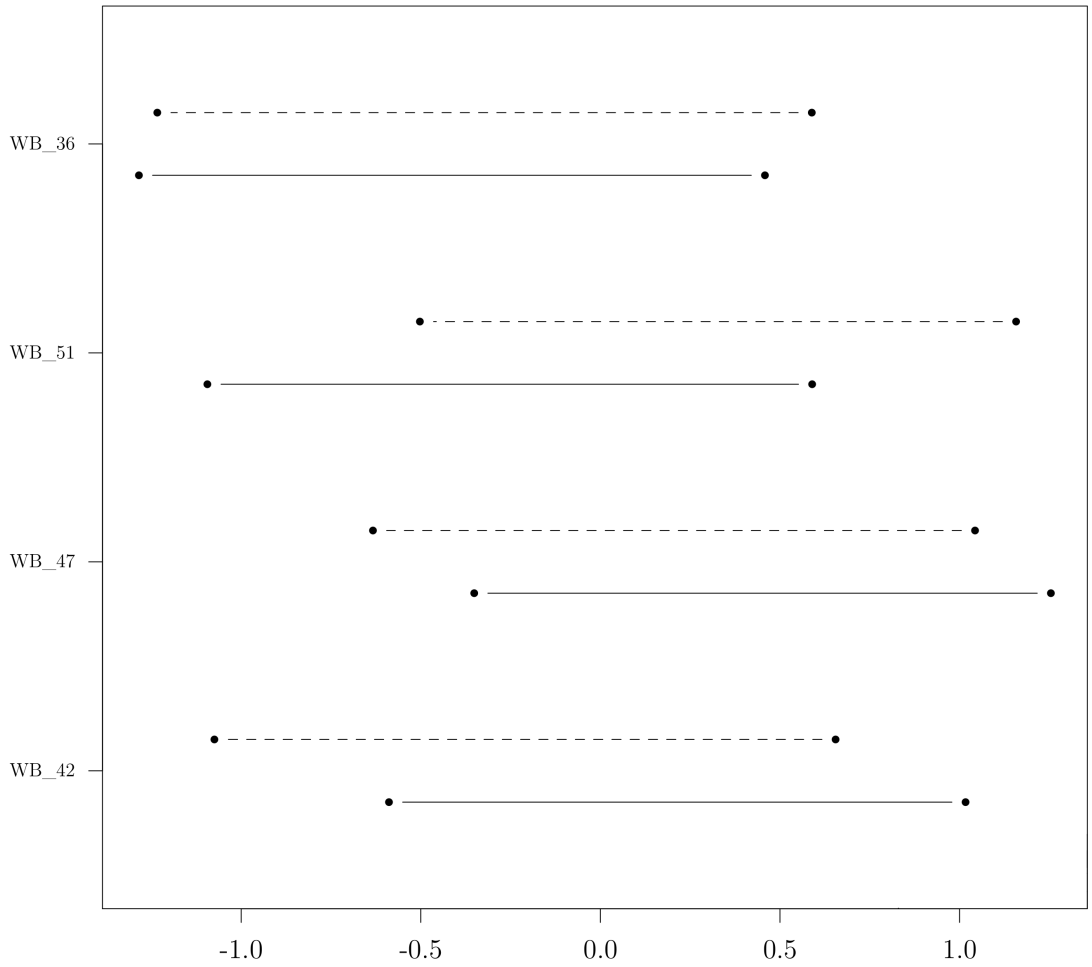

Abbildung 9.27: DIF-Plot. Itemschwierigkeiten aufgeteilt nach Geschlecht mit korrigierten Konfidenzintervall; Bonferroni-Korrektur. Skala Wortbil; MZP 3.

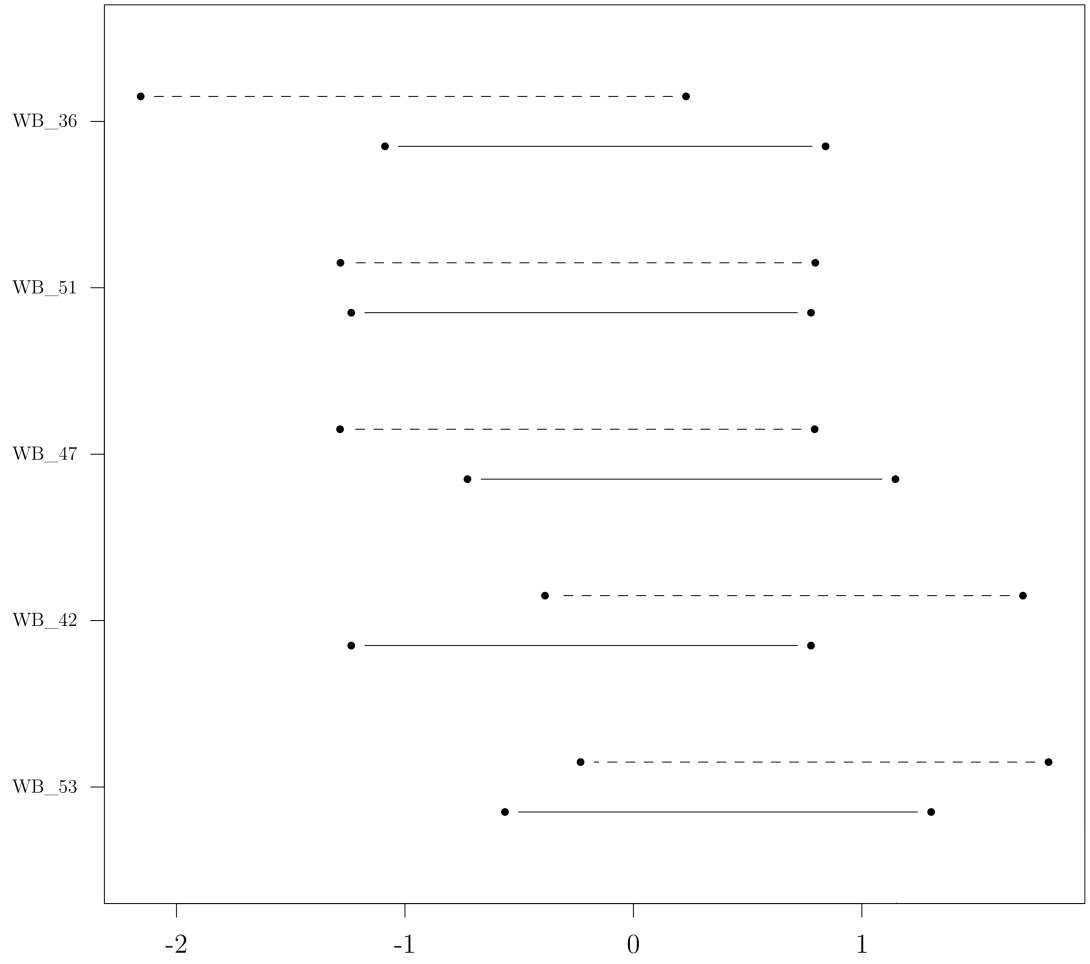

Abbildung 9.28: DIF-Plot. Itemschwierigkeiten aufgeteilt nach Geschlecht mit korrigierten Konfidenzintervall; Bonferroni-Korrektur. Skala Wortbil; MZP 4.

9.8.2.5 Modellvergleich

Der Modellvergleich der Skala Wortbil zwischen dem Rasch-Modell und dem Birnbaum-Modell (vgl. Tab. 9.26) bestätigt, dass die Daten zum dritten Messzeitpunkt besser durch das Rasch-Modell ($p < .05$) erklärt werden. Zum vierten Messzeitpunkt weisen sowohl ein niedriger AIC-Wert als auch BIC-Wert auf eine bessere Beschreibung der Daten durch das Birnbaum-Modell ($p < .001$) hin.

Tabelle 9.26: Modellvergleich - Skala Wortbil

Modell	MZP3		MZP4	
	AIC	BIC	AIC	BIC
1 PL	301.66	312.68	377.62	390.67
2 PL	296.58	314.22	357.48	379.22

Anmerkungen. 1 PL = Rasch-Modell; 2 PL = Birnbaum-Modell.

9.9 Beantwortung der Forschungsfragen

Forschungsfrage 1: Inwiefern kann auf Basis des sprachsystematischen Rechtschreibkompetenzmodells ein webbasiertes, zeit- und testökonomisches Instrument zur differenzierten Lernverlaufsdiagnostik im Primarbereich entwickelt werden?

Das in der vorliegenden Arbeit konstruierte und evaluierte Rechtschreibkompetenz-Messverfahren (ReKoMe) basiert auf der Theorie des sprachsystematischen Rechtschreibkompetenzmodells, ist zeit- und testökonomisch und im Unterricht webbasiert zur differenzierten Lernverlaufsdiagnostik einsetzbar.

Die schriftsprachtheoretische Fundierung des Messverfahrens ist zentral, weil daraus die Auswahl und Zuweisungen von Kategorien zur Analyse von Schreibprodukten erfolgt, die Rechtschreibkompetenz differentiell beschrieben werden kann und daraus unmittelbare Konsequenzen für anschließende Fördermaßnahmen resultieren (Naujokat, 2015). Der Itempool des Messverfahrens ist nach den fünf Teilkompetenzen des sprachsystematischen Rechtschreibkompetenzmodells eingeteilt und strukturiert. Jedes Testitem überprüft eine oder mehrere Teilkompetenzen, die gemäß Blatt et al. (2015) eindeutig spezifischen Wortteilen („Lupenstellen“) des sprachsystematischen Rechtschreibkompetenzmodells zugeordnet werden können. Die Grundmengen der Teilkompetenzen sind klar voneinander abgrenzbar und stellen homogene Teilmengen dar mit jeweils unterschiedlichen Anforderungsbereichen dar. Über die Onlinelernplattform Levumi ist das Messverfahren für Lehrkräfte und Schüler*innen einfach zugänglich und kann von den Schüler*innen eigenständig durchgeführt werden. Der Aufbau und die Bedienung des Messverfahrens sind kindgerecht, motivierend und einfach zu verstehen und anzuwenden. Durch den webbasierten Testmodus ist eine unabhängig von der Testleitung standardisierte Durchführungs-

und Auswertungsobjektivität gewährleistet, da die Antworten direkt eingegeben werden können. Die Ergebnisauswertung erfolgt unmittelbar automatisch, differenziert und steht nach Beendigung des Tests sofort zur Verfügung. Das Messverfahren generiert automatisch individuelle Lerngraphen für jede*n Schüler*in, die Informationen über Leistungsfortschritte, Leistungsrückschritte oder Stagnationen im Lernverlauf und Hinweise auf die Notwendigkeit für sofortiger Interventionsmaßnahmen liefern. Schüler*innen mit unterschiedlichen Kompetenzständen werden identifiziert und beschrieben, dadurch lassen sich gezielte Fördermaßnahmen und effiziente Unterrichtskonzepte ableiten. Die Durchführung einer Individualdiagnostik erfordert häufig eine Einzeltestung, die im Vergleich zu einer wesentlich ökonomischeren Gruppentestung mit einem erheblichen Mehraufwand verbunden ist (Brandt & Moosbrugger, 2020). Webbasierte diagnostische Tests ermöglichen, dass die Schüler*innen die Tests, die von der Lehrkraft unabhängige standardisierte Testinstruktionen enthalten, autonom in ihrem individuellen Tempo bearbeiten können und eine automatisierte Auswertung und Ergebnisdarstellung möglich ist (Gebhardt et al., 2016). Das Messverfahren kann sowohl für Einzeltests als auch für Gruppentests eingesetzt werden. Die Beantwortung der Testaufgaben erfolgt schriftlich per Tastatur durch die Schüler*innen. Dadurch entfällt die aufwendige und fehleranfällige manuelle Eingabe der Testergebnisse im Nachhinein. ReKoMe ist ein besonders zeitökonomisches Testverfahren. Mit wenigen Testaufgaben werden automatisch viele differenzierte Informationen zu den individuellen Rechtschreibkompetenzentwicklungen der Schüler*innen gewonnen. Die Verwendung des sprachsystematischen Rechtschreibkompetenzmodells bei der Operationalisierung der Items ermöglicht die Erfassung mehrerer Teilkompetenzen bzw. Merkmalsfacetten mit einer Testaufgabe.

Forschungsfrage 2: Inwiefern kann ein Algorithmus entwickelt werden, der die Testergebnisse auf Ganzwortebene und auf Ebene orthografischer Teilkompetenzen automatisiert analysiert, codiert und zuverlässige Ergebnisse liefert?

Die Auswertung von Schriftlösungen auf Basis des sprachsystematischen Rechtschreibkompetenzmodells ist komplex, zeitaufwendig und setzt ein hohes Fachwissen voraus. Der entwickelte und im ReKoMe implementierte Algorithmus codiert und analysiert zuverlässig automatisiert die Testergebnisse der Schüler*innen. Der Algorithmus basiert auf einem Kategoriensystem zur differenziellen Analyse der Schriftlösungen, anhand dessen auf Basis des sprachsystematischen Rechtschreibkompetenzmodells Struktureinheiten und Ausschlüsse der Wörter definiert werden. Grundlage für die Entwicklung des Algorithmus ist das Regelwerk zur computerbasierten Codierung von Wörtern auf Basis des sprachsystematischen Rechtschreibkompetenzmodells von Frahm (2013). Die entsprechenden Struktureinheiten der jeweiligen Teilkompetenzen werden im Kontext der gesamten Schriftlösung bewertet. Mit dem Algorithmus lassen sich auch Struktureinheiten identifizieren, die trotz korrekter Schreibungen der Struktureinheit als falsch zu werten sind, da vor oder nach der zu überprüfenden Struktureinheit falsche Buchstaben hinzugefügt wurden.

Der Vergleich einer manuellen Codierung der Testergebnisse mit einer automatisierten Codierung zeigt, dass mit dem Algorithmus die Schreiblösungen zu 99% richtig

analysiert und kategorisiert werden. Basis für die Überprüfung des Algorithmus bilden die im Rahmen der Paper-Pencil Studie zur Pilotierung der Testaufgaben erhobenen Daten. Es liegen insgesamt 833 unterschiedliche Schreiblösungen von insgesamt 54 Testitems vor. Für die unterschiedlichen Wörter liegen jeweils 2-69 verschiedene Schreibweisen vor wie z.B. für das Wort <Teller>: Tehler, Teler, Tella usw., die die Überprüfung der Korrektheit des Algorithmus ermöglichen.

Der Algorithmus prüft zuverlässig und automatisiert zum einem auf Wortebene, wie viele Wörter insgesamt richtig und wie viele Wörter falsch geschrieben werden. Zum anderen lässt sich mit dem Algorithmus auf Ebene der Teilkompetenzen des sprachsystematischen Kompetenzmodells die Art der Schreiblösungen innerhalb eines Wortes näher analysieren, um differenzierte Lernentwicklungsprofile erstellen zu können, anhand derer sich gezielte Förderimplikationen ableiten lassen. Durch die automatisierte Testauswertung, die im ReKoMe implementiert ist, eröffnen sich große Potenziale:

- Eine manuelle Codierung einer qualitativen Fehleranalyse ist aus ökonomischen Aspekten nicht praktikabel und erfordert grundlegende Kenntnisse der Sprachwissenschaften, die bei Lehrkräften nicht vorausgesetzt werden können (Frahm, 2013).
- Testreports und Rückmeldungen über die Testergebnisse können sofort nach Beendigung des Tests bereitgestellt und entsprechende Hinweise zur Förderung gegeben werden (Goldhammer & Kröhne, 2020).

Forschungsfrage 3: Ist das Rechtschreibkompetenz-Messverfahren (ReKoMe) ein reliables und valides Instrument zur differenzierten Lernverlaufsdiagnostik von Rechtschreibkompetenz in der dritten Grundschulklasse?

Mittels statistischer Analysen der klassischen Testtheorie und der Item-Response-Theorie konnte gezeigt werden, dass sich das Messverfahren insgesamt für den vorgesehenen Einsatzbereich in der dritten Klasse mit dem Fokus auf dem phonographisch-silbischen und morphologischen Prinzip sehr gut eignet, um Rechtschreibkompetenz im Lernverlauf sowohl auf Ebene des ganzen Wortes als auch auf Ebene der Teilkompetenzen valide zu messen. Es liegt eine im Mittel sehr gute Passung der einzelnen Skalen zum Rasch-Modell vor. Der gemittelte Wert des Infit MNSQ Werts liegt zwischen .89 und .99 und des Outfit MNSQ mit Werten zwischen .93 bis 1.14 nahe am Erwartungswert von 1 (Bond et al., 2020). Zudem differenzieren alle Items zwischen Personen mit unterschiedlichen Leistungsniveaus gut. Die Ergebnisse des bedingten Likelihood-Quotienten-Test (Andersen, 1973) bestätigen, dass für fast alle Skalen eine Itemhomogenität vorliegt und sich die Itemparameter zwischen zwei Gruppen, aufgeteilt nach dem Median, nicht signifikant unterscheiden. Für die Skala Wortbil ist die Durchführung eines Likelihood-Quotienten-Test zur Überprüfung der Itemhomogenität anhand des Teilungskriteriums Median aufgrund der Datenstruktur nicht möglich. Alle Skalen sind testfair, es liegt kein Differential Item Functioning vor. Beim Modellvergleich zwischen Rasch- und Birnbaum-Modell mittels Likelihood-Ratio-Test zeigt der BIC-Wert für alle Skalen - mit Ausnahme der Skala Wortbild

zum dritten Messzeitpunkt - eine bessere Passung zum Rasch-Modell. Der AIC-Wert deutet jedoch auf eine bessere Beschreibung der Daten durch das Birnbaum-Modell in den Skalen Quan, Morph, Peri und Wortbil hin. Der BIC-Wert ist im Vergleich zum AIC „strenger“ bezüglich der Modellkomplexität, berücksichtigt die Stichprobengröße und nimmt bei komplexeren Modellen mit mehr zu schätzenden Parametern höhere Werte an (Gäde et al., 2020) und ist für den Modellvergleich im Rahmen dieser Arbeit entscheidend. Mit der Skala Quan ist die Erfassung des Fähigkeitsniveaus von Schüler*innen Wörter auf Ganzwortebene orthografisch korrekt verschriftlichen zu können sehr gut (α: .93, Split-Half: r_{it} = .96; EAP Rel.: .89; WLE Rel.: .88) möglich. Es können zuverlässig differenzierte Aussagen über Personen mit unterschiedlichen Eigenschaftsausprägungen getroffen werden ($r_{it} = .42 - .43$). Die Skala misst sensitiv signifikante Leistungsfortschritte über zwei Messzeitpunkte. Die Item-Parameter der Skala Quan unterscheiden sich nicht signifikant aufgeteilt nach den Teilungskriterien Median und Geschlecht, es liegt kein Differential Item Functioning vor.

Die Teilkompetenz des Phonographisch-Silbischen Prinzip wird durch die Skala PhonSilb zuverlässig erfasst (EAP Rel.: .76-.79; WLE Rel.: .71-.73). Die Aufgaben scheinen zu den Messzeitpunkten mit 66% (MZP3) bis 70% (MZP4) durchschnittlich richtig gelöster Items eher leicht gewesen zu sein. Die durchschnittliche Trennschärfe liegt zu den Messzeitpunkten bei r_{it}= .38. Die Items eignen sich gut, um zwischen Personen mit unterschiedlichen Eigenschaftsausprägungen zu differenzieren.

Die Teilkompetenz des morphologischen Prinzips wird sehr zuverlässig durch die Items der Skala Morph erfasst (EAP Rel.: .80-.81; WLE Rel.: .74-.76) und differenziert gut zwischen den unterschiedlichen Leistungsniveaus ($r_{it} = .45 - .47$). Die Aufgaben liegen ebenfalls eher im leichten Bereich. Im Mittel wurden zwischen 58% (MZP3) und 64% (MZP4) der Items richtig gelöst. Das konnte erwartet werden, da dieses Rechtschreibprinzip im Lehrplan für die dritte Klasse vorgesehen ist.

Im Gegensatz zu den bereits beschriebenen Anforderungsniveaus der Skalen Quan, PhonSilb und Morph, welche die jeweiligen Teilkompetenzen des sprachsystematischen Rechtschreibkompetenzmodells repräsentieren, waren die in den Skalen Peri und Wortbil enthaltenen Items für die Schüler*innen durchschnittlich eher mittelschwer (Peri: M_3 = .50, M_4 = .54) bis schwer ($Wortbil$: M = .44). Während zum vierten Messzeitpunkt die Teilkompetenz des Peripheriebereichs durch die Items zuverlässig erfasst wird (α: .81, Split-Half:r_{it} = .92; $EAPRel.$: .71; $WLERel.$: .59) und gut zwischen den Leistungsniveaus differenziert (r_{it} = .46), zeigt sich für den dritten Messzeitpunkt eine differenziertere Befundslage. Der Wert des Cronbachs Alpha (.73) und die Werte der EAP- und WLE-Reliabilitäten (EAP Rel.: .67; WLE Rel.: .61) liegen im akzeptablen Bereich. Die Items eignen sich weniger gut, um zwischen leistungsstarken und weniger leistungsstarken Schüler*innenn zu trennen.

Für die Skala der Teilkompetenz des Wortbildungsprinzips konnte keine zufriedenstellende Evidenz bestätigt werden, da die Itemschwierigkeiten für die Stichprobe nicht im angemessenen Bereich lagen. Diese Ergebnisse müssen jedoch vorsichtig interpretiert werden, da sich im Vergleich zu den anderen Skalen weniger Items im

Itempool befinden (Lup = 5) und eine Teilkompetenz getestet wird, die in der dritten Klasse meistens noch nicht Gegenstand des Deutschunterrichts ist. Die Anzahl der Testitems hat Einfluss auf die Höhe des Cronbachs Alpha (Schermelleh-Engel & Gäde, 2020).

Forschungsfrage 4: Lässt sich die theoretisch postulierte mehrfaktorielle Struktur des sprachsystematischen Rechtschreibkompetenzmodells für die dritte Klasse empirisch nachweisen?

In den Studien IGLU-E Vorstudie, NEPS und HeLp wurde das sprachsystematische Rechtschreibkompetenzmodell auf seine Strukturen hin überprüft und mehrfach ein fünfstufiges Modell für die vierte und fünfte Klasse nachgewiesen, das die differentiellen Teilkompetenzen abbildet (Blatt, Prosch & Lorenz, 2016; Jarsinski, 2014; Naujokat, 2015; Voss et al., 2007). Für die dritte Klasse wurde dies bisher noch nicht geprüft. Inwiefern sich das dem Rechtschreibkompetenz-Messverfahren zugrunde gelegte theoretische Konstrukt bzw. die einzelnen Teilkompetenzen [16] des sprachsystematischen Rechtschreibkompetenzmodells (Blatt et al., 2015) in den vorliegenden Daten empirisch belegen lassen, wurde mit dem Vergleich eines einfaktoriellen und eines vierfaktoriellen Modells mittels einer konfirmatorischen Faktorenanalyse hinsichtlich ihrer Güte geprüft. Vergleicht man die beiden Modelle mit Hilfe des Akaike (AIC) und des Bayesian Information Criterion (BIC) als deskriptive Gütemaße, so weist das vierfaktorielle Modell mit geringeren Werten für AIC und BIC auf eine bessere Passung zwischen dem theoretisch zugrunde gelegten Konstrukt und den Daten hin. Dieses Ergebnis steht im Einklang mit dem Rahmenmodell des sprachsystematischen Rechtschreibkompetenzmodells, das die hier überprüften und bestätigten vier Faktoren postuliert [17].

9.10 Fallbeispiele

Die zuverlässige und präzise Bestimmung individueller Lernstände ist mit dem in dieser Arbeit konstruierten und evaluierten Rechtschreibkompetenz-Messverfahren (ReKoMe) möglich. Eine gezielte, stärkenorientierte Planung individueller Förderimplikationen lässt sich durch die automatische Erstellung von Kompetenzprofilen leicht umsetzen sowie Fehleinschätzungen in der Leistungsbeurteilung von Schüler*innen verhindern und Lernentwicklungen gezielt fördern. Dies wird im Folgenden an zwei ausgewählten Fallbeispielen verdeutlicht (Mau et al., 2018).

Die Schüler*innenprofile (vgl. Abb. 9.30, 9.32, 9.29, 9.31) zeigen die im Rahmen der Evaluationsstudie (vgl. Kap. 9) erhobene Rechtschreibkompetenzentwicklung zu fünf Messzeitpunkten über einen Zeitraum von einem halben Schuljahr. Zum Vergleich und zur Einordnung der Ergebnisse des ReKoMe liegen für beide Schülerinnen eine Leistungsmessung

[16] *P*honologisch-Silbisches Prinzip, Morphologisches Prinzip, Wortbildungs Prinzip, Peripheriebereich

[17] Die Wortübergreifende Teilkompetenz prüft die Groß- und Kleinschreibung, wurde in der vorliegenden Arbeit bei der Testauswertung auf Grund des webbasierten Testmodus nicht mit einbezogen.

mit der Hamburger Schreib-Probe sowie eine Leistungseinschätzung durch die unterrichtenden Lehrer*innen im Deutschunterricht vor [18] (vgl. Kap. 9.2). Die Schüler*innen haben jeweils einen diagnostizierten Förderbedarf in Deutsch und einen Migrationshintergrund. Die Schülerin aus dem ersten Beispiel wiederholt die dritte Klasse. Die Rechtschreibkompetenzen der Schüler*innen wurden durch die Lehrkraft auf einer Skala von eins bis sechs in Schulnoten eingeschätzt und jeweils als mangelhaft bewertet.

Die Individualgraphen (vgl. Abb. 9.29, 9.31) zeigen auf Wortebene den Lernstand der jeweiligen Schülerin pro Testung im Bezug zur Gesamtstichprobe der Evaluationsstudie. Die Prinzipienauswertung (vgl. Abb. 9.30, 9.32) zeigt auf Ebene der Teilkompetenzen des sprachsystematischen Rechtschreibkompetenzmodells den Lernstand in einem Balkendiagramm zum fünften Messzeitpunkt. Der orangefarbene Balken zeigt den prozentualen Anteil der im aktuellen Test richtig beantworteten Lupenstellen an. Der blaue Balken stellt die prozentual richtig beantworteten Lupenstellen der letzten Testung dar. Zusätzlich zeigt die Auswertung, wie viele Lupenstellen insgesamt und wie viele davon falsch beantwortet wurden. Der graue Balken stellt den durchschnittlichen Leistungsstand der Gesamtstichprobe der Evaluationsstudie in den einzelnen Teilkompetenzen dar.

Fallbeispiel 1

Die Leistung der Schülerin des ersten Beispiels ist auf Wortebene im Vergleich zur Gesamtstichprobe sowohl zum ersten als auch zum fünften Messzeitpunkt unterdurchschnittlich (vgl. Abb. 9.29).

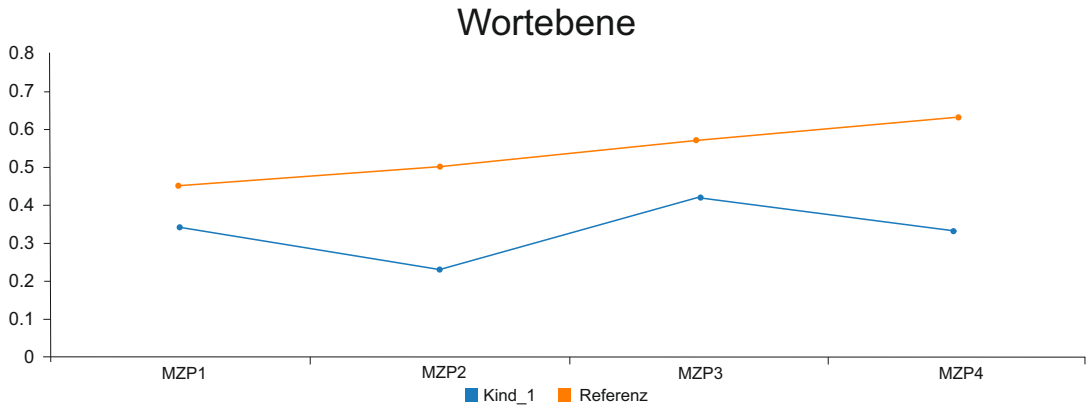

Abbildung 9.29: Individualgraph auf Wortebene Fallbeispiel 1

[18] Zur Überprüfung der externen Validität des ReKoMe wurde zum fünften Messzeitpunkt die Rechtschreibkompetenz der Schüler*innen mit der Hamburger Schreib-Probe erhoben. Zudem wurden die Lehrkräfte gebeten, die Rechtschreibkompetenz ihrer Schüler*innen in Schulnoten auf einer Skala von eins bis sechs einzuschätzen.

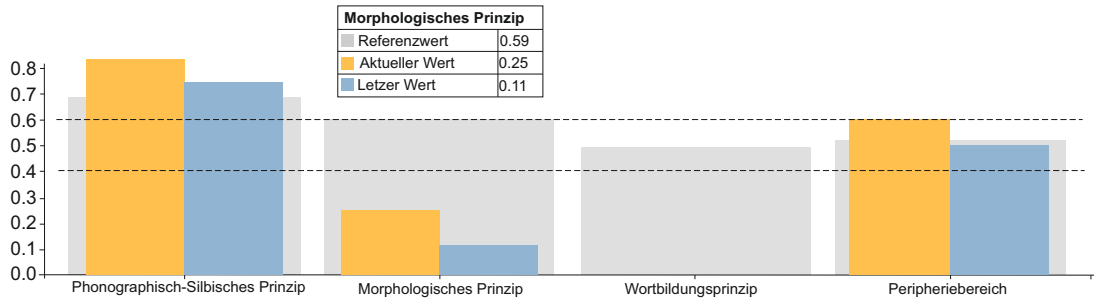

Abbildung 9.30: Prinzipienauswertung Fallbeispiel 1

Die Ergebnisse (vgl. Abb. 9.30) der jeweiligen Teilkompetenzen zeigen jedoch, dass lediglich die Ergebnisse in der Teilkompetenz des morphologischen Prinzips deutlich unterdurchschnittlich sind. Die Ergebnisse der Schülerin in der HSP (T = 50) sind durchschnittlich. Die Die Rechtschreibleistung der Schülerin wird von der Deutschlehkraft als mangelhaft bewertet (Mau et al., 2018).

Fallbeispiel 2

Im Gegensatz zu der von der Lehrkraft als mangelhaft eingeschätzten Rechtschreibkompetenz der Schülerin liegen die mit ReKoMe erhobenen Leistungen der zweiten Schülerin (Kind 2) auf Wortebene im Vergleich zur Gesamtstichprobe im überdurchschnittlichen Bereich (vgl. Abb. 9.31).

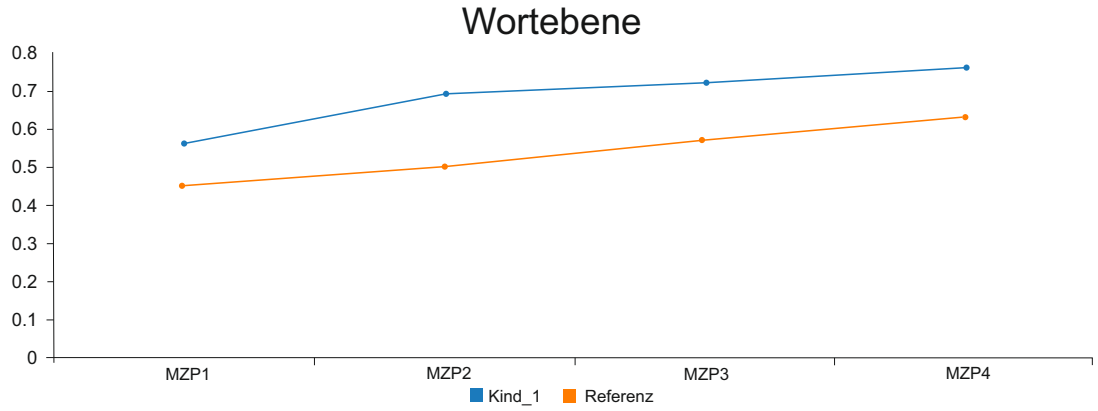

Abbildung 9.31: Individualgraph auf Wortebene Fallbeispiel 2

Die überdurchschnittliche Leistung in der Rechtschreibkompetenz zeigt sich zum fünften Messzeitpunkt auch auf der Ebene der Skalen der jeweiligen Teilkompetenzen des phonographisch-silbischen Prinzips, des morphologischen Prinzips, des Wortbildungsprinzip und im Peripheriebereich (vgl. Abb. 9.32) (Mau et al., 2018).

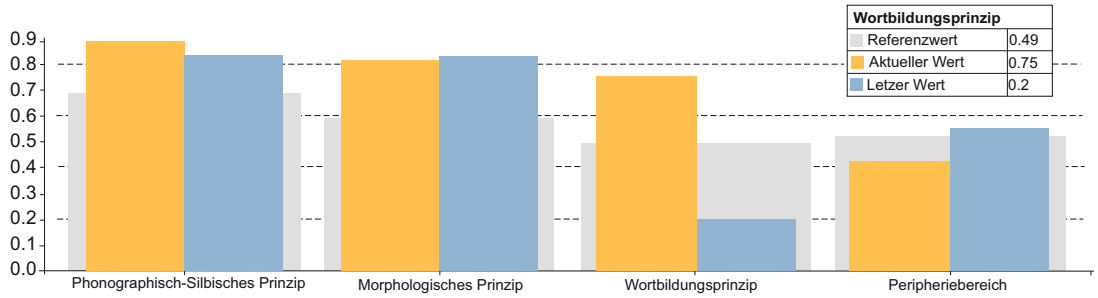

Abbildung 9.32: Prinzipienauswertung Fallbeispiel 2

Diese Ergebnisse werden auch durch das überdurchschnittliche Abschneiden in der Hamburger Schreib-Probe (HSP) bestätigt. [19]

Zusammenfassung

Wie die Fallbeispiele zeigen, kann ReKoMe einen entscheidenden Beitrag zum Abbau von Disparitäten im Bildungserfolg leisten. Die starken Fehleinschätzungen der Leistungen der Schüler*innen im Bereich Deutsch durch die Lehrkräfte können zum einen durch eine mangelnde diagnostische Kompetenz der Lehrer*innen erklärt werden, zum anderen aber auch durch eine generell geringere Leistungserwartung an die Schüler*innen aufgrund des Migrationshintergrundes. Dies kann weitreichende Konsequenzen für den Bildungsverlauf und insbesondere für die Schulempfehlung für die weiterführende Schule haben (Bonefeld et al., 2017; Tobisch et al., 2020). Die Abweichungen zwischen den Ergebnissen des ReKoMe, der HSP und der Lehrerbeurteilungen verdeutlichen die Notwendigkeit des Einsatzes des ReKoMe im Unterricht zur validen Leistungseinschätzung für die Schulpraxis.

Die Analyse der Ergebnisse der ersten Schülerin auf Subskalenebene zeigt, dass der Test vielfältige Möglichkeiten für eine individuelle Rechtschreibdiagnostik und -förderung bietet (Mau et al., 2018). Es ist wahrscheinlich, dass die Leistung der Schülerin auf Wortebene durch eine gezielte Förderung der korrekten Anwendung des morphologischen Prinzips verbessert werden könnte.

[19] In der HSP erreicht die Schülerin ein überdurchschnittliches Ergebnis (T=63).

10 Diskussion und Ausblick

Das Rechtschreibkompetenz-Messverfahren (ReKoMe) [1] zur differenzierten Lernverlaufsdiagnostik wurde unter Berücksichtigung orthografietheoretischer, fachdidaktischer, empirisch-methodischer, testökonomischer und zeitökonomischer Aspekte in der vorliegenden Arbeit konstruiert und evaluiert. Zentrales Anliegen ist die Verbindung der aus großen Schulleistungsstudien (z.B. NEPS, Pisa) gewonnenen Erkenntnisse zur sprachsystematischen Modellierung von Rechtschreibkompetenz mit den pädagogischen und didaktischen Interessen und Bedarfen von Lehrkräften und Schüler*innen.

Zunächst fand die Entwicklung und Konstruktion des Messverfahrens auf Basis des sprachsystematischen Rechtschreibkompetenzmodells und die Entwicklung des Algorithmus für eine theoretisch fundierte und automatisierte Ergebnisanalyse statt. Im Anschluss daran stand die Konzipierung des webbasierten Testdesigns sowie die Implementation auf der Onlinelernplattform Levumi im Vordergrund.

Mittels statistischer Analysen der klassischen Testtheorie und Item-Response-Theorie erfolgte die Pilotierung und Evaluation des webbasierten Prototypens, des Algorithmus und des Messverfahrens. Grundlage der Studie ist eine umfangreiche Stichprobe von 312 Schüler*innen aus 17 Grundschulklassen in drei Bundesländern. Einzelfallbeschreibungen zeigen, dass der Einsatz des Messverfahrens im entscheidenden Maße dazu beitragen kann, die Validität pädagogischer Entscheidungen im Deutschunterricht zu steigern und Bildungsdisparitäten abzubauen.

Die Diskussion der wichtigsten Ergebnisse erfolgt im nächsten Abschnitt. Anhand des Abschnitts 10.2 findet eine Darstellung der Grenzen der Untersuchung statt. Im letzten Abschnitt 10.3 werden die Perspektiven für die weitere Forschung aufgezeigt.

10.1 Diskussion der Ergebnisse

Das Messverfahren ist ein objektives, reliables und valides diagnostisches Instrument und erfüllt die in der Wissenschaft postulierten Qualitätskriterien, die für die Entwicklung eines Instruments zur Lernverlaufsdiagnostik zentral sind. Damit liegt erstmals ein webbasiertes und praxiszugängliches Instrument für die Primarstufe vor, das auf dem sprachsystematischen Kompetenzmodell basiert. Es ist geeignet, um Lehrkräfte bei der

[1] Das Rechtschreibkompetenz-Messverfahren (ReKoMe) ist seit 2017 auf der Onlinelernplattform www.Levumi.de bisher unter dem Namen „Wortdiktat“ (Mau, 2017) implementiert.

anspruchsvollen Aufgabe zu unterstützen, die individuellen Lernstände und Lernverläufe beim Rechtschreibkompetenzerwerb ihrer Schüler*innen zuverlässig zu diagnostizieren und engmaschig zu begleiten sowie Lernlücken frühzeitig zu identifizieren und passende Fördermaßnahmen einzuleiten.

Den sprachlichen Voraussetzungen und heterogenen Lernausgangslagen der Schüler*innen wird durch die Verwendung eines adäquaten und validierten Modells der Rechtschreibkompetenz sowie durch die Anwendung aktueller Ergebnisse der Sprach- und Schriftlichkeitsforschung Rechnung getragen.

Die Bedeutung einer webbasierten Diagnostik hinsichtlich einer ökonomischen Durchführung, einer automatisierten Auswertung und der Akzeptanz für den Einsatz von Instrumenten zur Lernverlaufsdiagnostik im Schulalltag ist beim Konstruktionsprozess berücksichtigt worden.

Die mit dem Messverfahren aus der Kompetenzmessung gewonnenen Informationen zu Lernständen und Lernverläufen werden durch die Implementierung auf der Onlineplattform Levumi für Akteur*innen im Bildungswesen nutzbar gemacht und liefern der Bildungsforschung wichtige Erkenntnisse zur Entwicklung der Rechtschreibkompetenz.

Sprachsystematische Modellierung von Rechtschreibkompetenz

Dem Messverfahren liegt eine theoriegeleitete Modellierung von Rechtschreibkompetenz auf Basis der aktuellen Forschungsergebnisse der Schriftlichkeitsforschung, der Fachdidaktik und der empirischen Bildungsforschung zugrunde. Dabei wurden curriculare Anforderungen, fachspezifische Zusammenhänge und kognitionspsychologische Modelle zur Rechtschreibkompetenz mit konkreten Testaufgaben verbunden und kohärent gestaltet. Die Abwendung einer auf Ebene des ganzen Wortes und auf Ebene von einzelnen Rechtschreibphänomenen angelegter Leistungsmessung - ohne die Einbettung in ein validiertes, adäquates Rechtschreibkompetenzmodell - hin zu einer kompetenzbasierten und fundierten Lernverlaufsdiagnostik im Bereich Rechtschreibung ist damit geebnet. Dies verhindert ein „empirisches Sortieren“ von Aufgabensammlungen bei der Testkonstruktion (Klieme & Leutner, 2006).

Die schriftsprachtheoretische Fundierung des Messinstruments ist zentral, weil daraus die Auswahl und Zuweisungen von Kategorien zur Analyse von Schreibprodukten erfolgt, die Rechtschreibkompetenz differentiell beschrieben werden kann und daraus unmittelbare Konsequenzen für anschließende Fördermaßnahmen resultieren (Naujokat, 2015). Die theoretische Fundierung vorhandener computergestützter Instrumente zur Lernverlaufsdiagnostik im Bereich Rechtschreibung basiert auf einem normbasierten und statischen Verständnis zum Schriftspracherwerb, indem die lautorientierte Schreibung als notwendige Entwicklungsstufe auf dem Weg zu einer orthografischen Schreibung angesehen wird. Das dem Messverfahren zugrunde liegende struktur- und prozessorientierte, sprachsystematische Rechtschreibkompetenzmodell eröffnet hingegen im Spannungsfeld zwischen generalisierenden Modellen und individualisierter Betrachtung der Rechtschreibkompetenzentwicklung große Potenziale für eine individualisierte Lernverlaufsdiagnostik (Bulut,

2018). Das Modell berücksichtigt die sprachlichen Voraussetzungen der Kinder und setzt anstelle der lautgetreuen Schreibung oder des Regellernens und Übens das Erkunden und Verstehen der Schriftstrukturen als Grundlage des Rechtschreiberwerbs (Blatt et al., 2015). Damit sind große Potenziale für die Lernverlaufsdiagnostik verbunden, weil eine schriftsystematische Analyse von Schreibprodukten möglich ist, welche die tatsächlichen individuellen Schriftentwicklungen der Schüler*innen „fernab der tradierten Brille einer idealtypischen Entwicklung“ überprüft (Weinhold et al., 2020, S.28).

Zeit- und testökonomische Vorteile

Der Testaufgabenpool ist durch insgesamt fünf Teilkompetenzen des sprachsystematischen Rechtschreibkompetenzmodells operationalisiert, dies ermöglicht die zeitökonomische Erfassung mehrerer Teilbereiche der Rechtschreibkompetenz mit einer Testaufgabe. Die Aufgaben werden je Testung und je Schüler*in nach einem proportional-zufälligen Verfahren erzeugt. Dadurch ist es möglich, eine sehr große Anzahl unterschiedlicher Testversionen zu generieren.

Der webbasierte Testmodus und die automatisierte Testauswertung der Testergebnisse bieten große zeitökonomische und testökonomische Vorteile. Eine standardisierte, von der Testleitung unabhängige Durchführungs- und Auswertungsobjektivität ist gewährleistet. Der Aufbau und die Bedienung des Messverfahrens sind im Verständnis und in der Anwendung kindgerecht, motivierend, haben eine leichte Navigation und können von den Schüler*innen eigenständig absolviert werden. Das Messverfahren kann jederzeit im Unterricht eingesetzt werden, da individuelle Testungen nach einer grundlegenden Einführung leicht möglich sind. Dies ist Grundvoraussetzung für eine praxisorientierte, individuelle Rechtschreibdiagnose und -förderung.

Die Beantwortung der Testaufgaben erfolgt durch die Schüler*innen schriftlich per Tastatur, Ergebnisreports stehen in Echtzeit zur Verfügung. Dadurch entfällt die aufwendige und fehlerbehaftete manuelle Eingabe der Testergebnisse. Bisherige Tests sind zwar computergestützt, erfordern aber eine Lehrkraft, die die Testdurchführung anleitet, die Schriftlösungen anschließend in ein Programm überträgt und auswertet, um daraus Förderimplikationen ableiten zu können.

Verbesserung von Lehr- und Lernprozessen

Das Messverfahren liefert wichtige diagnostische Informationen über die individuelle Entwicklung der Rechtschreibkompetenz. Diese sind eine wichtige Grundlage für Gespräche zwischen Lehrer*innen und Schüler*innen über Lernziele und für die Anpassung des unterrichtlichen Handelns an die individuellen Lernausgangslagen sowie für die Förderung des selbstregulierten Lernens. Der implementierte Algorithmus automatisiert die Messung und Analyse von Lernständen und -entwicklungen und erstellt in Echtzeit individuelle Rechtschreibkompetenzprofile der Schüler*innen. Durch die qualitative Analyse der Schreiblösungen werden wichtige Erkenntnisse zu den Präkonzepten und Denkwegen der

Schüler*innen auf dem Weg zum kompetenten Schreiben sichtbar gemacht. Dies ist ein sehr wichtiger Ausgangspunkt für eine individualisierte Förderung (Corvacho del Toro, 2016). Wichtig zu erkennen ist, „...wo die Probleme in der Erfassung des Orthographiesystems genau liegen, d.h. den zu fördernden Bereich zu spezifizieren und dass anhand dessen Erklärungs- und Übungssätze formuliert werden können, die sachlich angemessen und lernförderlich sind“(Corvacho del Toro, 2016). Dies ermöglicht die Planung individueller im Leistungsniveau angemessener, stärkenorientierter Lern- und Förderangebote. Die Schüler*innen erhalten zusätzlich automatisiert, nach Testende ein bezugsnormorientiertes Feedback per Audiodatei eingespielt.

Erkenntnisgewinn für Bildungforschung und Fachdidaktik

Die automatisierte, valide Bestimmung von Rechtschreibkompetenzentwicklung mit dem Messverfahren auf Basis des sprachsystematischen Kompetenzmodells bietet viele Potenziale für die weitere Forschung. Erkenntnisse zum Prozess und Verlauf der Rechtschreibkompetenzentwicklung können umfassend und differenziert, objektiv, zeit- und testökonomisch erfasst, analysiert und der Bildungsforschung und Fachdidaktik zugänglich gemacht werden. Daraus lassen sich weitere Fördermaßnahmen und Unterrichtskonzepte entwickeln und evaluieren. Der Algorithmus stellt bisher ein Novum dar und kann der Forschung für eine ökonomische Datenauswertung zum Erkenntnisgewinn zu Rechtschreibkompetenzentwicklungen auf Grundlage des sprachsystematischen Rechtschreibkompetenzmodells dienen.

10.2 Grenzen der Untersuchung

Bei der Interpretation und Generalisierung der Studienergebnisse müssen jedoch auch spezifische Limitationen berücksichtigt werden.

Die Stichproben der durchgeführten Studien basieren auf einem Convenience Sample und sind nicht repräsentativ. Eine mögliche Über- oder Unterrepräsentation bestimmter Gruppen innerhalb der Stichproben kann die Validität der Ergebnisse zusätzlich beeinträchtigen. Eine weitere Einschränkung betrifft den Einfluss von Kontextfaktoren. Die Datenerhebung fand während des regulären Schulbetriebs statt, und die Rahmenbedingungen variierten in den Schulen erheblich. Diese unkontrollierten Einflussfaktoren lassen sich nicht quantifizieren oder korrigieren. Zudem sind keine Aussagen über Interventionen im Rechtschreibunterricht zwischen den Messzeitpunkten möglich. Hätten die Kinder z.B. nach einem Test gezielt an der Schreibung der Testwörter gearbeitet, würde dies die Ergebnisse beeinflussen. Es ist wichtig, diese Limitationen bei der Interpretation der Ergebnisse im Blick zu behalten.

Die Ergebnisse der statistischen Analysen auf Basis der klassischen Testtheorie und der Item-Response-Theorie der vorliegenden explorativen Studie zur längsschnittlichen Untersuchung über die Güte des Rechtschreibkompetenz-Messverfahrens bestätigen, dass sich

das Messverfahren insgesamt für den vorgesehenen Einsatzbereich in der dritten Klasse mit dem Fokus auf dem phonographisch-silbischen und morphologischen Prinzip sehr gut eignet, um Rechtschreibkompetenz im Lernverlauf sowohl auf Ebene des ganzen Wortes als auch auf Ebene der Teilkompetenzen valide zu messen. Es liegt eine im Mittel sehr gute Passung der einzelnen Skalen zum Raschmodell vor. Insgesamt liegen für die Evaluationsstudie von ReKoMe durchschnittlich 5135 beantwortete Testaufgaben vor. Hervorzuheben ist auch, dass der entwickelte Algorithmus mit einer Wahrscheinlichkeit von 99% die Schreiblösungen richtig quantitativ und qualitativ codiert, analysiert und auswertet. Damit ist die Grundlage für eine qualitative Fehleranalyse geschaffen, anhand derer zuverlässig, schnell, effektiv und ressourcenorientiert Förderimplikationen abgeleitet werden können.

Die Durchführung einer Faktorenanalyse hatte zum Ziel, die theoretisch postulierte Struktur des sprachsystematischen Rechtschreibkompetenzmodells in den erhobenen Daten zu validieren. Obwohl die Analyse für ein vierfaktorielles Modell eine bessere Passung ergab als für ein einfaktorielles Modell, sind die Ergebnisse insgesamt als nicht zufriedenstellend zu bewerten. Dies wirft Fragen bezüglich der theoretischen Konzeptionalisierung und Operationalisierung der Testaufgaben auf. Ein möglicher Erklärungsansatz für die unbefriedigenden Ergebnisse könnte in der ungleichen Repräsentation der verschiedenen Teilkompetenzen im Itempool liegen. Zum Beispiel wurden für die phonographisch-silbische Teilkompetenz zum vierten Messzeitpunkt 1521 Testantworten erfasst, während für die Teilkompetenz der Wortbildung lediglich 173 Testantworten vorliegen. Dieses Ungleichgewicht könnte signifikante Auswirkungen auf die Validität der Modellannahmen und somit auf die Interpretation der Faktorenanalyse haben.

Im Vergleich zu den Skalen Quan, PhonSilb, Morph und Peri, welche die jeweiligen Teilkompetenzen mit den enthaltenen Testaufgaben valide erfassen, konnte für die Skala Wortbil keine zufriedenstellende Evidenz bestätigt werden. Die Aufgabenschwierigkeiten der Items lagen für die Stichprobe nicht im angemessenen Bereich. Diese Ergebnisse müssen jedoch vorsichtig interpretiert werden, da sich im Vergleich zu den anderen Skalen weniger Items im Itempool befinden und eine Teilkompetenz getestet wird, die in der dritten Klasse meistens noch nicht Gegenstand des Deutschunterrichts ist. Die Anzahl der Testitems hat Einfluss auf die Höhe des Cronbachs Alpha (Schermelleh-Engel & Gäde, 2020). Zur Bekräftigung der vielversprechenden Ergebnisse sollten weitere Testaufgaben für die Teilkompetenz des Wortbildungsprinzips entwickelt werden, damit alle Teilbereiche der globalen Rechtschreibkompetenz mit dem Messverfahren im Lernverlauf erfasst und auch für diesen Bereich detaillierte rechtschreibdidaktische Implikationen abgeleitet werden können. Die WLE- und EAP-Reliabilitäten für die Skalen Quan, Phon-Silb und Morph liegen hauptsächlich im guten Bereich und für die Skala Peri im zufriedenstellenden Bereich. Zur Erfassung des gesamten Leistungsspektrums von Personenfähigkeiten sollte der Itempool zukünftig um Items ergänzt werden, die auch die Randbereiche des Leistungsspektrums detailliert erfassen können. Bei der Überprüfung der Itemhomogenität traten bei einzelnen Items psychometrische Hürden auf, da diese keine passenden Antwortmuster für die Analyseverfahren aufwiesen und nicht mit dem LR-Test geschätzt werden konnten. Diese Items sollten an einer weiteren Stichprobe hinsichtlich der Homogenität geprüft werden.

10.3 Perspektiven für weitere Forschung

In der vorliegenden Studie wurde das Messverfahren erfolgreich evaluiert, allerdings sind die Ergebnisse statistisch nicht repräsentativ. Deshalb muss die endgültige Version an einer repräsentativen Stichprobe normiert werden, um geltende Testwertverteilungen und Normtabellen zu ermitteln. Um den Einsatzbereich des Messverfahrens zu erweitern, könnten weitere Aufgaben entwickelt und der Einsatz des webbasierten Testdesigns in den weiteren Klassenstufen der Primarstufe evaluiert werden. Die Auswertung der Ergebnisse einer durchgeführten Modeeffekt-Studie zu drei Messzeitpunkten mit Grundschüler*innen der dritten Klasse (N = 190) ist noch ausstehend. Anhand dessen können zukünftig Aussagen darüber getroffen werden, ob der webbasierte Testmodus einen Einfluss auf die Testergebnisse des Messverfahrens hat. Mit dem Messverfahren ist bisher die Erfassung der basalen Fähigkeit, auf Wortebene orthografisch korrekt schreiben zu können, möglich. Die Entwicklung weiterer Tests, die auch höhere kognitive Prozesse erfassen, könnten wichtige Hinweise zu Schwierigkeiten bei der Planung, der Verschriftlichung und dem Überarbeiten von adressatengerechten und sprachformalen korrekten Texten liefern (Winkes & Schaller, 2022). Eine zukünftige Studie, die weitere Faktoren des komplexen Wirkungsgeflechts des Lehr-Lerngeschehens beim Schriftspracherwerb wie z.B. die Lehrerkompetenz, die Unterrichtsmethode, die schulischen Rahmenbedingungen oder der sprachliche und soziale Hintergrund der Schüler*innen berücksichtigt, könnte wichtige Hinweise liefern, inwiefern sich diese unterrichtlichen Einflüsse in den Testergebnissen zur Validierung der Items niederschlagen. Die positiven Effekte der Lernverlaufsdiagnostik können sich nur dann einstellen, wenn ein entsprechender Wissensstand bei den Lehrkräften vorhanden ist und sie den fachgerechten Umgang mit den aus diagnostischen Verfahren gewonnen Informationen beherrschen. Darüber hinaus ist eine weitere Gelingensbedingung für formative Diagnostik, dass Lehrkräfte passende Förderimplikationen zur Verfügung gestellt bekommen. Im Kontext der Weiterentwicklung des Messverfahrens für die Schulpraxis spielen daher die Entwicklung und Evaluation geeigneter Fördermaterialien, die Fortbildung von Lehrkräften im Umgang mit dem Messverfahren sowie die Entwicklung adaptiver Tests eine wichtige Rolle.

Wichtige Potenziale zur Verbesserung der Rechtschreibkompetenz von Schüler*innen sind jetzt für die Diagnostik, Förderung, Unterricht und für den Erkenntnisgewinn über individuelle Rechtschreibkompetenzentwicklungen für Bildungsforschung und Fachdidaktik durch das Messverfahren verfügbar. Das Rechtschreibkompetenz-Messverfahren (ReKoMe) schließt eine bestehende Forschungslücke und dient sowohl der Grundlagenforschung als auch der praxisorientierten Anwendung.

11 Anhang

11.1 Tastaturschulung

Die heterogenen Vorerfahrung im Umgang mit einem Computer werden berücksichtigt, indem eine Tastaturschulung entwickelt wurde, die die Schüler*innen durch Erklärungen, Instruktionen und Aufgaben zur Bedienung mit der Tastatur vertraut macht. Diese sollte vor dem ersten Wortdiktat einmal erfolgt sein und kann bei Bedarf mehrmals wiederholt werden. Die Durchführungsdauer der Tastaturschulung beträgt ca. 10 Minuten. Die Tastaturschulung ist wie folgt aufgebaut:

Der kleine Drache Levumi stellt sich vor und erklärt das Ziel der folgenden Übungen:

„Hallo ich bin Levumi. Ich möchte heute mit dir üben, auf der Tastatur von einem Computer zu schreiben.“

Danach erfolgt via Kopfhörer die Aufforderung eine beliebige Taste zu drücken, um mit dem Programm zu starten. Die Übungen können so individuell angefangen werden.

> *„Bitte höre mir ganz genau zu, damit du weisst, was du tun sollst.“ „Bitte drücke jetzt irgendeine Taste auf der Tastatur, damit wir beginnen können.“* Als nächstes erfolgt die Einführung zweier Symbole, die im Verlauf des Tests wiederholt erscheinen:
>
> *„Hier siehst du auf dem Bild einen Kopfhörer. Immer, wenn du den Kopfhörer siehst, musst du gut zuhören, damit du genau weißt, was du tun sollst.“ „Das Bild mit der Tastatur zeigt dir, wann du etwas mit der Tastatur eingeben sollst.“*

In der ersten Aufgabe werden die Umlaute thematisiert. Diese sollen auf der Tastatur lokalisiert und im Folgenden in drei Beispielsätzen die fehlenden Umlaute ergänzt werden. Nach der Bearbeitung ertönt Levumi *Lob...* und es wird automatisch zur nächsten Übung weitergeleitet. Bei Fehlern gibt es die Möglichkeit, sich zu verbessern.

> *„ä, ü, ö sind Umlaute, bestimmt kennst du diese Buchstaben. Auf der Tastatur findest du diese Buchstaben auf der rechten Seite. Auf dem Bild siehst du, wo du die Buchstaben auf der Tastatur finden kannst.“ „Bitte drücke jetzt das ä, ü oder ö“ „In den folgenden Wörtern fehlen die Umlaute. Kannst du mir bitte helfen, diese wieder einzufügen? Bitte ergänze jetzt für mich die fehlenden Umlaute.“*

In der zweiten Aufgabe wird die Funktion der Löschtaste erläutert:

> *„Bestimmt hast du schon einmal an einem Computer etwas geschrieben, das du wieder löschen wolltest. Dafür benutzt man die Löschtaste. Auf dem Bild kannst du erkennen, wo du die Taste finden kannst. (Bild von Tastatur einblenden). Bitte drücke jetzt die Löschtaste auf der Tastatur.“*
> *„Jetzt benötige ich unbedingt deine Hilfe. Ich habe einen Satz geschrieben und möchte diesen wieder löschen. Bitte benutze die Löschtaste und lösche alle Buchstaben für mich.“*

Es folgt eine weitere Übungen zum ß:

> *„Nun geht es um einen Buchstaben, den es nur im Deutschen gibt, das ß. Es hat sich auf der Computertastatur oben rechts ziemlich gut versteckt. Drücke bitte die Taste für das ß auf der Tastatur.“* Nachdem das ß auf der Tastatur lokalisiert und gedrückt wurde, folgt eine Übung, in der das ß in Wörter einfügt werden soll: *„Das hast du gut gemacht! Bitte ergänze jetzt das fehlende ß in den Wörtern.“*

Die Großschreibung wird in der vierten Aufgabe behandelt:

> *„Um ein Wort großzuschreiben, muss man zwei Tasten gleichzeitig drücken: Den Pfeil, der nach oben zeigt, ganz links auf der Tastatur und eine Buchstabentaste. Am besten benutzt du dafür deine beiden Zeigefinger“.*
>
> Dann die Anfangsbuchstaben in drei Wörtern ergänzen. SchülerInnen darauf hinweisen, dass es sich um Großbuchstaben handelt. *„Bitte ergänze in den folgenden Wörtern die Großbuchstaben.“*

Die Tastaturschulung endet mit einer Abtippaufgabe. Die zu schreibenden Wörter sind gelb hinterlegt und müssen so schnell getippt werden, wie es geht. Nach Beendigung des Wortes muss die Leertaste gedrückt werden, um zum Folgewort zu kommen. Auf dem Bildschirm ist ein Zeitbalken zu sehen. Nach Ablauf von drei Minuten, wird die Übung automatisch beendet. *„Super, du hast es geschafft! Das hast du toll gemacht!“*

> *„Danke, dass du so gut mitmachst!“ „Jetzt hast du schon gut geübt, auf der Tastatur zu schreiben. Nun diktiere ich dir einen Satz. Bitte schreibe die Wörter so schnell du kannst. Aber achte genau auf die richtige Schreibung! Die Wörter, die du schreiben sollst, sind gelb makiert. Wenn du das Wort geschrieben hast, drücke die Leertaste, damit du zum nächsten Wort gelangst.“*

Während der Tastaturschulung erfolgen bei der richtigen Bearbeitung der Übung immer wieder visuelle Belohnungen in Form eines jubelnden Drachens (Levumi) mit einem Pokal.

Hat ein Kind Probleme bei der Erprobung einzelner Tasten, wird ein Bild mit der Tastenposition eingeblendet und eine Erklärung per Audiodatei eingespielt.

11.2 Itemanalysen

Tabelle 11.1: Itemschwierigkeiten, Infit- und Outfit-Koeffizienten der Skala Quan

Item	P_{i3}	P_{i4}	InfitMSQ3	OutfitMSQ3	InfitMSQ4	OutfitMSQ4
Kino	-3.81	-3.48	1.16	2.21	.95	.36
Beine	-3.57	-3.09	.95	.68	1.13	1.23
einer	-3.08	-2.13	.76	.91	.99	3.8
See	-2.77	-3.64	1.02	.64	.91	.77
Leine	-2.53	-2.45	1.02	1.46	.9	2.21
über	-2.43	-2.48	.86	.89	1	.62
Tee	-2.23	-2.76	1.04	.8	1.15	3.1
dort	-1.98	-1.89	1.14	2.06	1.14	1.04
Kilo	-1.61	-1.47	1.3	2.61	1.45	2.6
Zug	-1.51	-1.94	.96	.9	.86	.57
Flug	-1.26	-1.79	.89	.84	1.23	1.28
geht	-1.05	-1.12	.88	.79	.96	.86
Satz	-1.04	-1.63	1.15	1.52	.69	.52
Meer	-1.02	-1.15	1.14	1.32	.95	.82
spüren	-1.01	-1.19	1.27	1.85	1.16	1.91
sind	-.97	-.72	1.01	1.2	.87	.67
Geld	-.87	-1.08	.89	.97	.77	.59
Feld	-.72	-.22	1.02	.98	1.16	1.11
spülen	-.55	-.51	1.13	1.15	1.32	1.61
Platz	-.54	-1.15	.93	.85	.85	.84
Korb	-.53	-.48	1.02	1	1.11	1.01
seht	-.41	-.25	.87	.77	.87	.69
Teller	-.37	-.6	.93	.94	.85	.79
backt	-.21	-.3	.78	.65	.77	.85
Mäuse	-.2	-.47	1.06	1.56	.99	.95
kommt	-.19	-.58	.95	1.11	1.05	.91
Decke	-.09	-.35	.79	.7	.78	.62
Keller	0	0	.77	.68	.9	1.09
packt	0	.46	.89	.82	.86	.81
Läuse	.22	.61	.84	.73	.99	.98
Kräne	.31	-.07	1.24	1.72	1.4	5.75
Sieb	.35	.66	.98	.92	1.08	1.08
fliegen	.38	.38	.99	.99	.93	.93
Truhe	.44	.22	1.2	1.33	1.21	1.23
Zecke	.5	.49	.8	.73	.88	.76
winzig	.62	.72	1.03	1.01	1.18	1.05
sonnt	.64	.22	1.2	1.24	1.03	.96
Sahne	.66	.92	.9	.82	.85	.73
Bäche	.97	1.1	1.17	1.08	1.26	1.19
Stärke	1.07	1.41	.92	.9	.92	.84
Moor	1.2	.83	1.08	.92	.82	.65
Versteck	1.3	1.7	.9	.81	.61	.45
Video	1.37	1.62	.9	.83	.91	.69
Fahne	1.38	1.41	.98	.86	.96	.88
stärken	1.53	1.54	.87	.84	.91	.75
Strahl	1.71	1.83	.85	.6	.65	.45
Verdeck	1.86	1.93	.72	.55	.61	.42
bissig	2.33	2.46	1	.87	1.24	1.48
fließen	2.33	3.09	.95	.64	.99	.56
Weihnachtsbaum	2.51	2.72	1.02	4.15	.7	.43
Strähne	3.13	3.42	.85	.77	1.16	1.41
Nationalmannschaft	4.33	5.42	.73	.28	.46	.07
Adventskranz	5.52	4.26	.92	.26	.87	.28

Anmerkungen. P_{i3} = Itemschwierigkeiten zum dritten Messzeitpunkt, P_{i4} = Itemschwierigkeiten zum vierten Messzeitpunkt; MSQ = Mean Square Fit

Tabelle 11.2: Itemschwierigkeiten, Infit- und Outfit-Koeffizienten der Skala PhonSilb

Item	P_{i3}	P_{i4}	InfitMSQ3	OutfitMSQ3	InfitMSQ4	OutfitMSQ4
Kino	-2.67	-2.52	1.16	2.00	.86	.39
Läuse	-2.61	-1.74	.77	2.29	.90	.91
Beine	-2.42	-2.14	.94	.76	1.03	.88
Sahne	-2.11	-2.55	1.05	1.56	1.08	1.25
einer	-1.95	-1.19	.79	.47	.96	3.24
Mäuse	-1.56	-3.36	.98	1.21	.97	1.07
Leine	-1.41	-1.42	.85	.68	.81	.60
Strahl	-1.40	-1.00	.92	.57	.75	.43
über	-1.35	-1.47	.79	.78	.91	.55
Kräne	-1.02	-.64	1.25	1.64	.83	.75
Bäche	-.68	-.89	.84	.80	.91	.79
Fahne	-.61	-1.08	1.23	1.58	1.24	1.65
Strähne	-.61	-.11	.88	.98	.79	.72
Kilo	-.57	-.52	1.16	1.70	1.39	1.66
spüren	-.42	-.71	.90	1.04	.96	.68
Stärke	-.07	.21	1.02	1.02	.95	.99
Keller	.00	.00	.76	.69	.86	.86
spülen	.19	-.10	.89	.91	1.04	1.01
stärken	.27	.16	.98	1.00	.90	.81
Teller	.59	.34	.92	1.04	.95	.77
winzig	.77	.74	1.02	.92	1.04	.96
Decke	.84	.60	.80	.72	.84	.72
fliegen	1.27	1.33	.89	.85	.92	.95
Truhe	1.35	1.12	1.03	.97	1.22	1.28
Zecke	1.36	1.44	.84	.77	.79	.71
bissig	2.18	2.43	1.08	1.16	.95	.90
fließen	3.09	3.65	.87	.59	.82	.55
Adventskranz	4.24	4.69	1.12	3.70	.82	.58
Nationalmannschaft	4.43	4.14	.65	.48	.87	1.38

Anmerkungen. P_{i3} = Itemschwierigkeiten zum dritten Messzeitpunkt, P_{i4} = Itemschwierigkeiten zum vierten Messzeitpunkt; MSQ = Mean Square Fit

Tabelle 11.3: Itemschwierigkeiten, Infit- und Outfit-Koeffizienten der Skala Morph

Item	P_{i3}	P_{i4}	InfitMSQ3	OutfitMSQ3	InfitMSQ4	OutfitMSQ4
spüren	-2.01	-1.73	1.09	1.36	1.2	1.9
spülen	-1.79	-1.61	1.14	1.01	1.21	1.61
Zug	-1.18	-1.53	.96	.97	.8	.52
Flug	-.86	-1.33	.86	.73	1.17	1.03
Satz	-.67	-1.19	1.05	.99	.7	.77
geht	-.66	-.7	.88	.85	1.04	.91
Geld	-.49	-.65	.8	.88	.75	.61
Feld	-.34	.21	1.03	1.04	1.07	1.11
Platz	-.15	-.71	.9	.79	.85	.72
seht	-.02	.21	.92	.83	.85	.72
Korb	0	0	1	1.01	1.04	.96
Mäuse	.02	-.04	1.07	1.24	.95	.87
kommt	.14	-.17	.94	.94	1.04	.94
backt	.19	.11	.81	.69	.75	.71
Kräne	.24	-.06	1.27	1.49	1.37	2.23
packt	.43	.87	.83	.74	.93	.92
Läuse	.51	1.06	.85	.73	.92	.88
Adventskranz	.62	.83	1.23	1.3	1.31	1.41
Sieb	.69	1.06	.88	.78	.97	1.01
Weihnachtsbaum	.76	.94	1.3	1.63	1.04	1.01
sonnt	.99	.68	1.16	1.22	.98	.98
Versteck	1	1.37	.84	.75	.69	.57
Bäche	1.18	1.28	1.14	1.12	1.27	1.18
Verdeck	1.53	1.15	.84	.75	.75	.63

Anmerkungen. P_{i3} = Itemschwierigkeiten zum dritten Messzeitpunkt, P_{i4} = Itemschwierigkeiten zum vierten Messzeitpunkt; MSQ = Mean Square Fit

Tabelle 11.4: Itemschwierigkeiten, Infit- und Outfit-Koeffizienten der Skala Peri

Item	P_{i3}	P_{i4}	InfitMSQ3	OutfitMSQ3	InfitMSQ4	OutfitMSQ4
See	-3.02	-3.83	0.67	0.32	0.55	0.19
Tee	-2.56	-3.02	0.7	0.38	0.95	9.77
dort	-2.15	-2.28	1.04	1.17	1.07	1.23
Meer	-1.24	-1.42	1.1	1.48	0.82	0.7
sind	-1.13	-0.91	1.04	1.89	0.9	0.85
Fahne	0	0	0.85	0.66	0.81	0.57
Sahne	0.4	0.76	0.82	0.73	0.81	0.78
Moor	1.01	0.65	0.95	0.81	0.88	0.86
Video	1.26	1.59	0.87	1.09	0.87	0.61
Strahl	1.52	1.87	0.86	0.78	0.63	0.4
Weihnachtsbaum	1.81	1.61	0.94	1.55	1.16	1.41
Nationalmannschaft	3.01	3.66	0.83	0.43	0.69	0.3

Anmerkungen. P_{i3} = Itemschwierigkeiten zum dritten Messzeitpunkt, P_{i4} = Itemschwierigkeiten zum vierten Messzeitpunkt; MSQ = Mean Square Fit

Tabelle 11.5: Itemschwierigkeiten, Infit- und Outfit-Koeffizienten der Skala Wortbildung

Item	P_{i3}	P_{i4}	InfitMSQ3	OutfitMSQ3	InfitMSQ4	OutfitMSQ4
winzig	-.58	-.46	.97	1.02	.85	.76
Versteck	-.08	.24	.9	.85	.81	.78
Adventskranz	-.08	-.33	1.3	1.37	1.11	1.21
Verdeck	0	0	.68	.66	1.03	1.03
bissig	.17	.01	1.06	1.09	1.12	1.13

Anmerkungen. P_{i3} = Itemschwierigkeiten zum dritten Messzeitpunkt, P_{i4} = Itemschwierigkeiten zum vierten Messzeitpunkt; MSQ = Mean Square Fit

Literatur

Albert, J. O. (2017). *Entwicklung und Analyse eines Algorithmus zur Kategorisierung von Rechtschreibfehlern* (Bachelorarbeit). Christian-Albrechts-Universität zu Kiel. Kiel, Deutschland.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123–140.

Bangel, M. & Müller, A. (2018). Strukturorientiertes Rechtschreiblernen. Ergebnis einer Interventionsstudie zur Wortschreibung in Klasse 5 mit Blick auf schwache Lerner/-innen. *Didaktik Deutsch: Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur*, *23*(45), 29–49.

Bastian, M. (2017). *Entwicklung eines Schülerzugangs für die Plattform LeVuMi am Beispiel eines Diktiertests* (Bachelorarbeit). Christian-Albrechts-Universität zu Kiel. Kiel, Deutschland.

Becker, T. (2008). Modelle des Schriftspracherwerbs: Eine kritische Bestandsaufnahme. *Didaktik Deutsch: Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur*, (25), 78–95.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, *5*(1), 7–74.

Black, P. & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Blatt, I., Prosch, A., Jarsinski, S., Bos, W. & Kander, M. (2021). *Entwicklung der Rechtschreibkompetenz von Klassenstufe 5 bis 9* (LIfBi Working Paper Nr. 101). Leibniz-Institut für Bildungsverläufe. https://doi.org/10.5157/LIfBi:WP101:1.0

Blatt, I., Voss, A., Kowalski, K. & Jarsinski, S. (2015). Messung von Rechtschreibleistung und empirische Kompetenzmodellierung. In U. Bredel & T. Reißig (Hrsg.), *Weiterführender Orthographieunterricht* (2. korr. Aufl., S. 226–256). Schneider Verlag.

Blatt, I. (2010). Sprachsystematische Rechtschreibdidaktik: Konzept, Materialien, Tests. In U. Bredel, A. Müller & G. Hinney (Hrsg.), *Schriftsystem und Schrifterwerb* (S. 101–132). De Gruyter. https://doi.org/10.1515/9783110232257.101

Blatt, I. & Frahm, S. (2013). Explorative Analysen zur Entwicklung der Rechtschreibkompetenz im Rahmen der NEPS-Studie (Klassenstufe 5-7). *Didaktik Deutsch: Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur*, *18*(34), 13–36.

Blatt, I. & Pagel, B. (2009). Die interaktive Tafel als Medium im sprachlichen Anfangsunterricht. *Grundschulunterricht*, *1*, 25–29.

Blatt, I. & Prosch, A. (2016). Rechtschreibkompetenz in der Sekundarstufe I – Ausgewählte Ergebnisse aus der Längsschnittstudie Nationales Bildungspanel (NEPS). In M. Krelle & M. Senn (Hrsg.), *Qualitäten von Deutschunterricht.* (S. 85–108).

Blatt, I., Prosch, A. & Frahm, S. (2016). Erfassung der Rechtschreibkompetenz in der Rechtschreibstudie „Nationales Bildungspanel“. Studiendesign und Ergebnisse. In B. Mesch & C. Noack (Hrsg.), *System, Norm und Gebrauch – Orthographische Kompetenz im Performanz im Spannungsfeld zwischen System, Norm und Empirie* (S. 53–72).

Blatt, I., Prosch, A. & Lorenz, C. (2016). Erhebung der Rechtschreibkompetenz am Ende der Grundschulzeit: Ausgewählte Ergebnisse aus einer Großpilotstudie im Rahmen des Nationalen Bildungspanels. *Zeitschrift für Grundschulforschung*, *9*(2), 125–138.

Blumenthal, S. (2022). Lernverlaufsdiagnostik. In M. Gebhardt, D. Scheer & M. Schurig (Hrsg.), *Handbuch der sonderpädagogischen Diagnostik. Grundlagen und Konzepte der Statusdiagnostik, Prozessdiagnostik und Förderplanung* (S. 633–648). Universitätsbibliothek. https://doi.org/10.5283/epub.53149

Blumenthal, S., Blumenthal, Y., Ehrich, K. & Mahlau, K. (2020). Die „Lernlinie Rechtschreibung“ - ein Diagnostik- und Fördertool für die Grundschule. *Weitblick. Magazin für Lehrkräfte an Gemeinschafts- und Förderschulen*, (2), 22–29.

Blumenthal, S., Sikora, S. & Mahlau, K. (2021). Lernverlaufsdiagnostik im Rechtschreibunterricht der Grundschule: Konstruktion und Güte eines curriculumbasierten Messverfahrens. *Diagnostica*, *67*(2), 49–61. https://doi.org/10.1026/0012-1924/a000261

Böhme, K., Engelbert, M. & Weirich, S. (2017). Beschreibung der im Fach Deutsch untersuchten Kompetenzen. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich & N. Haag (Hrsg.), *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich* (S. 20–30). Waxmann.

Bond, T., Yan, Z. & Heene, M. (Hrsg.). (2020). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (4th). Routledge. https://doi.org/10.4324/9780429030499

Bonefeld, M., Dickhäuser, O., Janke, S., Praetorius, A.-K. & Dresel, M. (2017). Migrationsbedingte Disparitäten in der Notenvergabe nach dem Übergang auf das Gymnasium. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie*, (1), 11–23.

Brandt, H. & Moosbrugger, H. (2020). Planungsaspekte und Konstruktionsphasen von Tests und Fragebogen. *Testtheorie und Fragebogenkonstruktion* (S. 39–66). Springer. https://doi.org/10.1007/978-3-662-61532-4_3

Bredel, U., Fuhrhop, N. & Noack, C. (2017). *Wie Kinder lesen und schreiben lernen* (2. Aufl.). Francke Verlag.

Bremerich-Vos, A. (2009). Das „Mimesisbild der Alphabetschrift“ und didaktische Kontroversen zum Schriftspracherwerb. In E. Birk & J. G. Schneider (Hrsg.), *Philosophie der Schrift* (S. 43–58). Max Niemeyer Verlag.

Brügelmann, H. & Brinkmann, E. (1994). Stufen des Schriftspracherwerbs und Ansätze zu seiner Förderung. In H. Brügelmann & S. Richter (Hrsg.), *Wie Kinder recht schreiben lernen* (S. 44–52). Schneider Verlag Hohengehren.

Brügelmann, H. & Brinkmann, E. (2013). Rechtschreibung, Rechtschreibförderung, Rechtschreiblernen im Anfangsunterricht. Anmerkungen zu dem Statement von Ursula Bredel für die Anhörung vor dem Schulausschuss der Hamburgischen Bürgerschaft am 3.12.2013. https://grundschulverband.de/wp-content/uploads/2017/02/bredel-schulausschuss.pdf

Budde, M., Riegler, S. & Wiprächtiger-Geppert, M. (2012). *Sprachdidaktik*. Akademie Verlag.

Bulut, N. (2018). *Individuelle Rechtschreibentwicklung: Eine Längsschnittuntersuchung zur Bedeutung von Einflussfaktoren auf die Wortschreibung*. Schneider Verlag Hohengehren.

Bulut, N. (2019). Rechtschreibentwicklung messen. In I. Kaplan & I. Petersen (Hrsg.), *Schreibkompetenzen messen, beurteilen und fördern* (S. 39–56). Waxmann Verlag.

Bußmann, H. (2008). *Lexikon der Sprachwissenschaft* (4., durchgesehene und bibliografisch ergänzte Auflage). Alfred Kröner Verlag.

Capizzi, A. M. & Fuchs, L. S. (2005). Effects of curriculum-based measurement with and without diagnostic feedback on teacher planning. *Remedial and Special Education*, *26*(3), 159–174. https://doi.org/10.1177/07419325050260030401

Chomsky, N. (1968). *Language and Mind* (1st). Harcourt, Brace & World.

Corvacho del Toro, I. (2016). Zur qualitativen Rechtschreibfehleranalyse und einer schriftsystematischen lernförderlichen Behandlung der Rechtschreibstörung. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie*, *44*(5), 397–408.

Corvacho del Toro, I. & Günther, T. (2013). Zum Effekt des Fachwissens von Lehrkräften auf die Rechtschreibleistung von Grundschülern. *Lernen und Lernstörungen*, *2*(1), 21–33.

Coulmas, F. (1981). *Über Schrift* (Bd. 378). Suhrkamp.

Dehn, M. (1985). Über die sprachanalytische Tätigkeit des Kindes beim Schreibenlernen. *Diskussion Deutsch*, *16*(81), 25–51.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional children*, *52*(3), 219–232.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The journal of special education*, *37*(3), 184–192.

Deno, S. L. & Mirkin, P. K. (1977). *Data-based program modification: A manual*. Minnesota, MN: University of Minnesota.

Döring, N. & Bortz, J. (2016). Forschungsmethoden und evaluation. *Wiesbaden: Springerverlag*.

Dürscheid, C. (2016). *Einführung in die Schriftlinguistik. Mit einem Kapitel zur Typographie von Jürgen Spitzmüller* (5. aktual. u. korr. Aufl.). Vandenhoeck & Ruprecht.

Eichler, W. (1976). Zur linguistischen Fehleranalyse von Spontanschreibungen bei Vor- und Grundschulkindern. *A. Hofer (Hg.): Lesenlernen: Theorie und Unterricht.(1. Aufl.). Düsseldorf: Pädagogischer Verlag Schwann*, 246–264.

Eid, M. & Petermann, F. (2006). Aufgaben, Zielsetzungen und Strategien der Psychologischen Diagnostik. In F. Petermann & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 15–25). Hogrefe Verlag.

Eisenberg, P. (1989). Die Schreibsilbe im Deutschen. In P. Eisenberg & H. Günther (Hrsg.), *Schriftsystem und Orthographie* (S. 57–84). Max Niemeyer Verlag.

Eisenberg, P. (2016). Phonem und Graphem. Der Buchstabe und die Schriftstruktur des Wortes. *Dudenredaktion (Hg.): Duden. Die Grammatik. Berlin: Duden Verlag*, 19–60.

Eisenberg, P. (2020a). *Grundriss der deutschen Grammatik. Das Wort* (5. Aufl.). Springer. https://doi.org/10.1007/978-3-476-05096-0

Eisenberg, P. (2020b). *Grundriss der deutschen Grammatik. Der Satz* (5. Aufl.). Springer. https://doi.org/10.1007/978-3-476-05094-6

Eisenberg, P. & Fuhrhop, N. (2007). Schulgraphematik und Orthographie. *Zeitschrift für Sprachwissenschaft*, *26*, 15–41. https://doi.org/10.1515/ZFS.2007.004

Embretson, S. E. (1996). The new rules of measurement. *Psychological assessment*, *8*(4), 341–349.

Fay, J. (2013). Rechtschreiblernen in der Primarstufe. In S. Gailberger & F. Wietzke (Hrsg.), *Handbuch Kompetenzorientierter Deutschunterricht* (S. 172–194). Beltz.

Fay, J. & Berkling, K. (2013). Rechtschreibdiagnostik. In J. Fay (Hrsg.), *(Schrift-) Sprachdiagnostik heute* (S. 84–108). Schneider-Verlag.

Fay, J., Berkling, K. & Stüker, S. (2012). Automatische Analyse von Rechtschreibfähigkeit auf Basis von Speech-Processing-Technologien. *Didaktik Deutsch: Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur*, *17*(33), 14–36.

Felder, E. (2003). Sprache als Medium und Gegenstand des Unterrichts. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der Deutschen Sprache – ein Handbuch* (S. 42–51). Schöningh.

Fisseni, H. J. (2004). *Lehrbuch der psychologischen Diagnostik: mit Hinweisen zur Intervention*. Hogrefe Verlag.

Förster, N. & Souvignier, E. (2014). Learning progress assessment and goal setting: Effects on reading achievement, reading motivation and reading self-concept. *Learning and Instruction*, *32*, 91–100.

Förster, N. & Souvignier, E. (2015). Effects of providing teachers with information about their students' reading progress. *School Psychology Review*, *44*(1), 60–75.

Frahm, S. (2013). *Computerbasierte Testung der Rechtschreibleistung in Klasse fünf - eine empirische Studie zu Mode-Effekten im Kontext des Nationalen Bildungspanels*. Logos Verlag Berlin GmbH.

Frith, U. (1986). Psychologische Aspekte des orthographischen Wissens: Entwicklung und Entwicklungsstörung. *New trends in graphemics and orthography*, 218–233.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review*, *33*(2), 188–192.

Fuchs, L. S. & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional children*, *53*(3), 199–208.

Fuhrhop, N. (2020). *Orthografie* (5. akt. Aufl.). Universitätsverlag Winter.

Gäde, Schermelleh-Engel, K. & Brandt, H. (2020). Konfirmatorische Faktorenanalyse (CFA). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkon-*

struktion (3., vollst. neu bearb., erw. u. akt. Aufl., S. 615–659). Springer. https://doi.org/10.1007978-3-662-61532-4_24

Gebhardt, M., Diehl, K. & Mühling, A. (2016). Online Lernverlaufsmessung für alle SchülerInnen in inklusiven Klassen. www. LEVUMI. de. *Zeitschrift für Heilpädagogik*, *67*(10), 444–454.

Glück, H. & Rödel, M. (2016). *Metzler Lexikon. Sprache*. Metzler Verlag.

Goldhammer, F. & Kröhne, U. (2020). Computerbasiertes Assessment. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3., vollst. neu bearb., erw. u. akt. Aufl., S. 119–141). Springer. https://doi.org/10.1007/978-3-662-61532-4_6

Granzer, D., Böhme, K. & Köller, O. (2008). Kompetenzmodelle und Aufgabenentwicklung für die standardisierte Leistungsmessung im Fach Deutsch. *Lernstandsbestimmung im Fach Deutsch. Gute Aufgaben für den Unterricht* (S. 10–49).

Günther, H. (1988). *Schriftliche Sprache. Strukturen geschriebener Wörter und ihre Verarbeitung beim Lesen*. Niemeyer.

Günther, H. (2010). *Beiträge zur Didaktik der Schriftlichkeit* (Bd. 6). Gilles und Francke Verlag.

Hanke, P. & Schwippert, K. (2005). Orthographische Lernprozesse im Grundschulbereich. Ergebnisse aus Mehrebenenanalysen. *Unterrichtswissenschaft*, *33*(1), 70–91. https://doi.org/10.25656/01:5788

Hartig, J. (2008). Kompetenzen als Ergebnisse von Bildungsprozessen. *Kompetenzerfassung in pädagogischen Handlungsfeldern. Theorien, Konzepte und Methoden*, *26*, 15–26.

Hartig, J. & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. *Leistung und Leistungsdiagnostik* (S. 127–143). Springer.

Hasselhorn, M., Schneider, W. & Trautwein, U. (2014). *Lernverlaufsdiagnostik* (Bd. 12). Hogrefe Verlag.

Hein, C. & Blatt, I. (2016). Untersuchung von Unterrichtsbedingungen zu Erwerb und Entwicklung der Schriftkompetenz – Ausgewählte Ergebnisse einer Interventionsstudie mit Kontrollklassen von Klasse 1 bis 3. In M. Krelle & M. Senn (Hrsg.), *Qualitäten von Deutschunterricht* (S. 109–138). Fillibach bei Klett.

Heward, W. L. (2003). Ten faulty notions about teaching and learning that hinder the effectiveness of special education. *The journal of special education*, *36*(4), 186–205.

Hinney, G. (2014). Lesen- und Schreibenlernen mit der Silbe – ein sprachdidaktischer Fortschritt? In D. Wrobel & A. Müller (Hrsg.), *Bildungsmedien für den Deutschunterricht. Vielfalt - Entwicklungen - Herausforderungen.* (S. 143–169).

Hinney, G. (1997). *Neubestimmung von Lerninhalten für den Rechtschreibunterricht: Ein fachdidaktischer Beitrag zur Schriftaneignung als Problemlöseprozess*. Lang.

Hinney, G. (2004). Das silbenbasierte Rechtschreibwissensmodell als Grundlage für Lernsoftware. Konzeptionelle Überlegungen. In I. Blatt & W. Hartmann (Hrsg.), *Schreibprozesse im medialen Wandel* (S. 72–85). Schneider-Verlag.

Hinney, G. (2010). Wortschreibungskompetenz und sprachbewusster Unterricht. In U. Bredel, A. Müller & G. Hinney (Hrsg.), *Schriftsystem und Schrifterwerb* (S. 47–100). De Gruyter. https://doi.org/10.1515/9783110232257.47

Hinney, G. (2015). Was ist Rechtschreibkompetenz? In U. Bredel, T. Reißig & W. Ulrich (Hrsg.), *Weiterführender Orthographieerwerb* (S. 191–225). Schneider-Verlag.

Hinney, G. & Menzel, W. (1998). Didaktik des Rechtschreibens. In G. Lange, K. Neumann & W. Ziesenis (Hrsg.), *Taschenbuch des Deutschunterrichts. Grundfragen und Praxis der Sprach- und Literaturdidaktik* (6. vollständig überarbeitete Auflage, S. 258–304). Schneider Hohengehren.

Hintz, A.-M. & Grünke, M. (2009). Einschätzungen von angehenden Lehrkräften für Sonder- und allgemeine Schulen zur Wirksamkeit von Interventionen für den Schriftspracherwerb bei lernschwachen Kindern. *Empirische Sonderpädagogik*, *1*(1), 45–61. https://doi.org/10.25656/01:9463

Hosp, M. K. & Hosp, J. L. (2003). Curriculum-based measurement for reading, spelling, and math: How to do it and why. *Preventing School Failure: Alternative Education for Children and Youth*, *48*(1), 10–17.

Hosp, M. K., Hosp, J. L. & Howell, K. W. (2016). *The ABCs of CBM: A practical guide to curriculum-based measurement*. Guilford Publications.

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik*. Beltz Verlag. https://books.google.de/books?id=YFTOwAEACAAJ

Jagemann, S. & Weinhold, S. (2018). Schriftspracherwerb. *Empirische Forschung in der Deutschdidaktik*, *3*, 253–268.

Jarsinski, S. (2014). *Quantitative Datenanalyse zur längsschnittlichen Erfassung der Rechtschreibkompetenz in NEPS unter besonderer Berücksichtigung der Kompetenzstruktur und der Einflussfaktoren* (Diss.). TU Dortmund. https://doi.org/10.17877/DE290R-6602

Jungjohann, J., Mau, L., Diehl, K. & Gebhardt, M. (2019). *Levumi: Handbuch für Lehrkräfte Deutsch*. Technische Universität Dortmund. https://doi.org/10.17877/DE290R-20903

Kelava, A. & Moosbrugger, H. (2020a). Deskriptivstatistische Itemanalyse und Testwertbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3., vollst. neu bearb., erw. u. akt. Aufl., S. 143–158). Springer. https://doi.org/10.1007/978-3-662-61532-4_7

Kelava, A. & Moosbrugger, H. (2020b). Einführung in die Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3., vollst. neu bearb., erw. u. akt. Aufl., S. 369–409). Springer. https://doi.org/10.1007/978-3-662-61532-4_16

Kelava, A., Schermelleh-Engel, K. & Mayer, A. (2020). Latent-State-Trait-Theorie (LST-Theorie). *Testtheorie und Fragebogenkonstruktion* (S. 687–711). Springer.

Kerstan, T. (2022, 1. Juli). Grundschüler : "Die Lage ist wirklich besorgniserregend". *Zeit Online*. https://doi.org/https://www.zeit.de/gesellschaft/2022-07/grundschueler-schulleistung-bildung-pandemie-iqb-studie

Klauer, K. J. (2011). Lernverlaufsdiagnostik - Konzept, Schwierigkeiten und Möglichkeiten. *Empirische Sonderpädagogik*, *3*(3), 207–224.

Klauer, K. J. (2014). Formative Leistungsdiagnostik: Historischer Hintergrund und Weiterentwicklung zur Lernverlaufsdiagnostik. In M. Hasselhorn, U. Trautwein & W. Schneider (Hrsg.), *Lernverlaufsdiagnostik* (S. 1–18). Hogrefe.

Klicpera, C., Schabmann, A., Gasteiger-Klicpera, B. & Schmidt, B. (2020). *Legasthenie-LRS: Modelle, Diagnose, Therapie und Förderung*. Ernst Reinhardt Verlag.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E. et al. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. Bonn, Berlin: BMBF.

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik*, *52*(6), 876–903.

Klieme, E., Maag-Merki, K. & Hartig, J. (2007). Kompetenzbegriff und Bedeutung von Kompetenzen im Bildungswesen. *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik*, 5–15.

Koller, I., Alexandrowicz, R. & Hatzinger, R. (2012). *Das Rasch Modell in der Praxis: Eine Einführung in eRm*. utb GmbH.

Kruse, N. & Reichardt, A. (2016). *Wie viel Rechtschreibung brauchen Grundschulkinder?: Positionen und Perspektiven zum Rechtschreibunterricht in der Grundschule*. Erich Schmidt Verlag.

Kultusminister Konferenz. (2005). *Beschlüsse der Kultusministerkonferenz. Bildungsstandards im Fach Deutsch für den Primarbereich (Jahrgangsstufe 4)*. Sekretariat der Ständigen Konferenz der Kultusminister Konferenz der Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_10_15-Bildungsstandards-Deutsch-Primar.pdf

Langfeldt, H. P. & Imhof, M. (1999). Schulleistungsdiagnostik. In C. Perleth & A. Ziegler (Hrsg.), *Pädagogische Psychologie: Grundlagen und Anwendungsfelder* (S. 280–289). Hans Huber.

Leutner, D. (2013). Perspektiven pädagogischer Interventionsforschung. In E. Severing & R. Weiß (Hrsg.), *Qualitätsentwicklung in der Berufsbildungsforschung* (S. 17–28). W. Bertelsmann Verlag. https://doi.org/https://doi.org/10.3278/111-054w017

Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, *10*(8), 707–710.

Linacre, J. M. et al. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch measurement transactions*, *16*(2), 878.

List-Ivankovic, J. (2013). *Evaluation von Bildungsprojekten auf der Grundlage von Inventaren: Entwicklung und Erprobung eines Ansatzes im Rahmen von europäischen Projekten* (Dissertation). Georg-August-Universität Göttingen.

Maas, U. (2010). *Grundzüge der deutschen Orthographie* (Bd. 120). Walter de Gruyter.

Maas, U. (2013). Die deutsche Orthographie. *Die Rechtschreibung als Ausbau des sprachlichen Wissens. Arbeitsfassung*, *20*.

Maier, U. (2014). *Leistungsdiagnostik in Schule und Unterricht: Schülerleistungen messen, bewerten und fördern*. UTB.

Maier, U., Randler, C. & Wolf, N. (2016). Effekte von computergestützten, formativen Tests mit unterschiedlichen Rückmeldeformaten auf Lernleistungen im naturwissenschaftlichen Unterricht. *Zeitschrift für Pädagogik*, *62*(2), 241–262.

Mau, L., Mühling, A. & Diehl, K. (2018). Lernverlaufsmessung mit Levumi - Ein curriculumbasiertes Messverfahren für Rechtschreibung in der dritten Klasse. In T. Jungmann, B. Gierschner, M. Meindl & S. Sallat (Hrsg.), *Sprach- und Bildungshorizonte. Wahrnehmen - Beschreiben - Erweitern* (S. 180–185). Springer.

May, P. (2010). *Hamburger Schreib-Probe HSP 1-9: Diagnose orthografischer Kompetenz; zur Erfassung der grundlegenden Rechtschreibstrategien mit der Hamburger Schreibprobe; [Manual]*. Verlag für pädagogische Medien.

May, P. (2009). Auswertung der Rechtschreibleistung nach dem Strategiediagnosekonzept. *Kompetenzmodelle der Orthographie. Empirische Befunde und förderdiagnostische Möglichkeiten. Beiträge*, *10*, 75–89.

May, P. (2013). *HSP: Manual, Handbuch: Diagnose orthografischer Kompetenz zur Erfassung der grundlegenden Rechtschreibkompetenzen*. VPM, Verlag für pädagogische Medien, Klett.

May, P., Vielauf, U. & Malitzky, V. (2002). *HSP Hamburger Schreib-Probe: Diagnose orthographischer Kompetenz: zur Erfassung der grundlegenden Rechtschreibstrategien mit der Hamburger Schreibprobe*. Verlag für Pädagogische Medien.

Moosbrugger, H. & Kelava, A. (2020). Qualitätsanforderungen an Tests und Fragebogen („Gütekriterien“). *Testtheorie und Fragebogenkonstruktion* (S. 13–38). Springer.

Mühling, A., Gebhardt, M. & Diehl, K. (2017). Formative Diagnostik durch die Onlineplattform LEVUMI. *Informatik-Spektrum*, *40*, 556–561.

Naujokat, K. (2015). *Die Validierung von Rechtschreibkompetenzmodellen im Rahmen einer empirisch vergleichenden Analyse orthografischer Leistungstests* (Diss.). TU Dortmund. https://doi.org/10.17877/DE290R-16563

Naumann, C. L. (2015). Die Architektur der Schrift im Deutschen ist asymmetrisch. Was bedeutet das für den Erwerb? In N. Kruse & A. Reichardt (Hrsg.), *Wie viel Rechtschreibung brauchen Grundschüler? Kritische Bilanz und neue Perspektiven des Rechtschreibunterrichts in der Grundschule* (S. 67–79). Erich Schmidt Verlag.

Neef, M. (2011). *Die Graphematik des Deutschen*. De Gruyter. https://doi.org/10.1515/9783110914856

Nerius, D. (2007). *Deutsche Orthographie* (4., neubearb. Aufl.). Olms.

Noack, C. (2001). *Regularitäten der deutschen Orthographie und ihre Deregulierung: Eine computerbasierte diachrone Untersuchung zu ausgewählten Sonderbereichen der deutschen Rechtschreibung* (Dissertation). Universität Osnabrück.

Ossner, J. (2006). Kompetenzen und Kompetenzmodelle im Deutschunterricht. *Didaktik Deutsch*, *21*(2006), 5–19.

Ossner, J. (2018). » Bericht über 12 Jahre Rat für deutsche Rechtschreibung.«. *Didaktik Deutsch*, *23*(44), 101–111.

Parshall, C. G., Harmes, J., Davey, T. & Pashley, P. J. (2010). Innovative item types for computerized testing. In W. J. van der Linden & C. A. Glas (Hrsg.), *Elements of adaptive testing* (S. 215–230). Springer.

Paul, H. (1880). *Prinzipien der Sprachgeschichte*. Max Niemeyer.

Petermann, F. & Wirtz, M. (2023). *Psychologische Diagnostik im Dorsch Lexikon der Psychologie*. Verfügbar 1. April 2023 unter https://dorsch.hogrefe.com/gebiet/psychologische-diagnostik

Prenzel, M., Walter, O. & Frey, A. (2007). PISA misst Kompetenzen. *Psychologische Rundschau*, *58*(2), 128–136.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Rat für deutsche Rechtschreibung. (2018). *Deutsche Rechtschreibung. Regeln und Wörterverzeichnis: Aktualisierte Fassung des amtlichen Regelwerks entsprechend den Empfehlungen des Rats für deutsche Rechtschreibung 2016* (Techn. Ber.). Rat für deutsche Rechtschreibung. https://www.rechtschreibrat.com/DOX/rfdr_Regeln_2016_redigiert_2018.pdf

Reber, K. (2017). *Prävention von Lese- und Rechtschreibstörungen im Unterricht: Systematischer Schriftspracherwerb von Anfang an* (2., überarbeitete Auflage). Reinhardt Ernst.

Reber, K. & Kirch, M. (2013). Richtig schreiben lernen: Kompetenzorientierter, inklusiver Rechtschreibunterricht. *Praxis Sprache*, *4*, 254–257.

Reichardt, A. (2015). *Rechtschreibung im Textraum - Modellierungen der Schreibkompetenz in der Grundschule* (Bd. 9). Gilles & Francke.

Riehme, J. (1974). *Probleme und Methoden des Rechtschreibunterrichts*. Volk und Wissen Volkseigener Verlag.

Roos, J. & Schöler, H. (2009). *Entwicklung des Schriftspracherwerbs in der Grundschule*. Springer.

Rose, N. (2020). Parameterschätzung und Messgenauigkeit in der Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3., vollst. neu bearb., erw. u. akt. Aufl., S. 447–500). Springer. https://doi.org/10.1007978-3-662-61532-4_19

Rost, J. (2004). *Lehrbuch Testtheorie: Testkonstruktion (2. vollst. überarb. und erw. Aufl.)* Huber.

Saussure, F. d. (1916). Cours de Linguistique Générale. Paris: Payot. Dt. Übers.: Grundfragen der allg. *Sprachwissenschaft. Berlin: de Gruyter*.

Scheerer-Neumann, G. (2007). Rechtschreiben. In U. Moser & S. Knost (Hrsg.), *Sonderpädagogik des Lernens* (S. 539–568). Hogrefe.

Scheerer-Neumann, G. (2008). Der Erwerb der basalen Lese- und Schreibfähigkeiten. In H. Günther & O. Ludwig (Hrsg.), *Ein interdisziplinäres Handbuch internationaler Forschung* (S. 1153–1169). De Gruyter Mouton. https://doi.org/10.1515/9783110147445.2.8.1153

Schermelleh-Engel, K. & Gäde, J. (2020). Modellbasierte Methoden der Reliabilitätsschätzung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3., vollst. neu bearb., erw. u. akt. Aufl., S. 335–368). Springer. https://doi.org/10.1007978-3-662-61532-4_15

Schneider, W. (2008). Entwicklung und Erfassung der Rechtschreibkompetenz im Jugend- und Erwachsenenalter. In W. Schneider, H. Marx & M. Hasselhorn (Hrsg.), *Diagnostik von Rechtschreibleistungen und -kompetenz* (S. 145–157). Hogrefe Verlag.

Schröder, E. (2019). *Der Lerngegenstand Wortschreibung aus der Sicht von Lehrenden*. Springer.

Scriven, M. (1991). *Evaluation Thesaurus* (4th). Sage Publications.

Scriven, M. (1967). The Methodology of Evaluation. In R. Tyler, R. Gagné & M. Scriven (Hrsg.), *Perspectives of Curriculum Evaluation, AERA Monograph Series on Curriculum Evaluation* (S. 39–83). Rand McNally.

Shapiro, E. S., Edwards, L. & Zigmond, N. (2005). Progress monitoring of mathematics among students with learning disabilities. *Assessment for Effective Intervention*, *30*(2), 15–32.

Souvignier, E., Förster, N. & Schulte, E. (2014). Wirksamkeit formativen Assessments – Evaluation des Ansatzes der Lernverlaufsdiagnostik. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (S. 221–238). Hogrefe Verlag.

Spitta, G. (1988). Geben wir den Kindern Zeit, damit sie aus ihren Fehlern lernen können. *Die Grundschulzeitschrift*, *12*, 2–12.

Stanat, P., Schipolowski, S., Rjosk, C., Weirich, S. & Haag, N. (2017). *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich*. Waxmann.

Stanat, P., Schipolowski, S., Schneider, R., Sachse, K. A., Weirich, S. & Henschel, S. (2022). *IQB-Bildungstrend 2021. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im dritten Ländervergleich*. Waxmann Verlag. https://doi.org/10.31244/9783830996064

Stecker, P. M., Fuchs, L. S. & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, *42*(8), 795–819.

Strathmann, A., Klauer, K. J. & Greisbach, M. (2010). Lernverlaufdiagnostik - Dargestellt am Beispiel der Entwicklung der Rechtschreibkompetenz in der Grundschule. *Empirische Sonderpädagogik*, *2*(1), 64–77.

Strathmann, A. & Klauer, K. J. (2008). Diagnostik des Lernverlaufs. Eine Pilotstudie am Beispiel der Entwicklung der Rechtschreibkompetenz. *Sonderpädagogik*, *38*(1), 5–24.

Thomé, G. (2003). Entwicklung der basalen Rechtschreibkenntnisse. *Didaktik der deutschen Sprache. Ein Handbuch*, *1*, 369–380.

Thurstone, L. L. (1928). Attitudes can be measured. *American journal of Sociology*, *33*(4), 529–554.

Tobisch, A., Klapproth, F. & Dresel, M. (2020). Werden Kinder mit Migrationshintergrund durch Lehrkräfte benachteiligt? *In-Mind Magazin*, *2020*(3). https://de.in-mind.org/article/werden-kinder-mit-migrationshintergrund-durch-lehrkraefte-benachteiligt

Valtin, R. (2000). Ein Entwicklungsmodell des Rechtschreibenlernens. In R. Valtin (Hrsg.), *Rechtschreiben lernen in den Klassen 1–6: Grundlagen und didaktische Hilfen* (S. 17–22). Arbeitskreis Grundschule e.V.

Voß, S. & Hartke, B. (2014). Curriculumbasierte Messverfahren (CBM) als Methode der formativen Leistungsdiagnostik im RTI-Ansatz. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (S. 83–99). Hogrefe.

Voß, S., Blumenthal, Y., Ehrich, K. & Mahlau, K. (2020). Multimodale Diagnostik als Ausgangspunkt für spezifische Förderung. *Zeitschrift für Heilpädagogik*, *71*, 88–99.

Voß, S., Sikora, S. & Mahlau, K. (2017). Vorschlag zur Konzeption eines curriculumbasierten Messverfahrens zur Erfassung der Rechtschreibleistungen im Grundschulbereich. *Empirische Sonderpädagogik*, *9*(2), 184–194.

Voss, A., Blatt, I. & Kowalski, K. (2007). Zur Erfassung orthographischer Kompetenz in IGLU 2006: Dargestellt an einem sprachsystematischen Test auf Grundlage von Daten aus der IGLU-Voruntersuchung. *Didaktik Deutsch*, *23*(2007), 15–33.

Walter, J. (2014). Lernfortschrittsdiagnostik Lesen (LDL) und Verlaufsdiagnostik sinnerfassenden Lesens (VSL): Zwei Verfahren als Instrumente einer formativ orientierten Lesediagnostik. *Lernverlaufsdiagnostik*, 165–201.

Walter, J., Clausen-Suhr, K. & Clausen-Suhr, K. (2018). *LDO: Lernfortschrittsdiagnostik Orthographie: ein computergestütztes Verfahren zur längsschnittlichen Erfassung orthographischer Kompetenzen für Zweit- und Drittklässler*. Hogrefe.

Weinert, F. E. (2001). *Leistungsmessung in Schulen* (2. Aufl.). Beltz.

Weinhold, S., Jagemann, S. & Stahr, B. (2020). Entwicklungsmuster von (schwachen) Rechtschreibleistungen und individuellen Schriftlösungen. In I. Rautenberg (Hrsg.), *Evidenzbasierte Forschung zum Schriftspracherwerb* (S. 5–30). Schneider Verlag.

Wilbert, J. & Linnemann, M. (2011). Kriterien zur Analyse eines Tests zur Lernverlaufsdiagnostik. *Empirische Sonderpädagogik*, *3*(3), 225–242. https://doi.org/10.25656/01:9325

Winkes, J. & Schaller, P. (2022). Lernverlaufsdiagnostik Schreiben (LVD–Schreiben): Reliabilität, Validität und Sensitivität für mittelfristige Lernfortschritte im deutschsprachigen Raum. *Vierteljahresschrift für Heilpädagogik und ihre Nachbargebiete*, 1–26.