Dissertation

zur Erlangung des akademischen Grades eines

Doktors der Philosophie (Dr. phil.)

Quantitative Datenanalyse zur längsschnittlichen Erfassung der Rechtschreibkompetenz in NEPS unter besonderer Berücksichtigung der Kompetenzstruktur und der Einflussfaktoren

An der

Fakultät Erziehungswissenschaft,

Psychologie und Soziologie

der Technischen Universität Dortmund

Vorgelegt von

Dipl.-Päd. Stephan Jarsinski

Geboren am 08.05.1983 in Herdecke

bei

Prof. Dr. Wilfried Bos

Prof. Dr. (em.) Inge Blatt

Dortmund im Juni 2014

IFS Institut für Schulentwicklungs-forschung

technische universität dortmund

NEPS Nationales Bildungspanel

An dieser Stelle möchte ich mich bei all denjenigen bedanken, die mich während der Erstellung dieser Dissertation unterstützt und motiviert haben.

Mein Dank gilt Herrn Prof. Dr. Wilfried Bos und Frau Prof. Dr. (em.) Inge Blatt, die mit viel Geduld und Mühen meine Arbeit durch ihre Ideen, ihre Anregungen und ihre konstruktive Kritik bereicherten. Ebenfalls möchte ich meinen Kolleginnen und Kollegen der TU Dortmund und der Universität Hamburg danken, ohne die die Fertigstellung dieser Arbeit nicht möglich gewesen wäre.

Ein ganz besonderer Dank geht an meine Eltern für ihre Unterstützung und ihr Vertrauen, die die Grundsteine für meinen Weg gelegt haben. Abschließend danke ich den weiteren Menschen in meinem Leben, die in jeder Hinsicht für mich da sind.

– Diese Dissertation ist meinem Großvater gewidmet –

INHALTSVERZEICHNIS

ABKÜRZUNGSVERZEICHNIS

ES K5	Entwicklungsstudie in Klassenstufe 5
ES K6	Entwicklungsstudie in Klassenstufe 6
ES K7	Entwicklungsstudie in Klassenstufe 7
GP K8	Großpilot in Klassenstufe 8
GP K9	Großpilot in Klassenstufe 9
HE K5	Haupterhebung in Klassenstufe 5
HE K7	Haupterhebung in Klassenstufe 7
IRT	Item-Response-Theory
SRT	Sprachsystematischer Rechtschreibtest
GW	Ganzes Wort
PSP	Phonographisch-silbisches Prinzip
MP	Morphologisches Prinzip
PB	Peripheriebereich
PDW	Prinzip der Wortbildung
WUP	Wortübergreifendes Prinzip
MW	Mittelwert
SD	Standardabweichung
SE	Standardfehler
MIN	Minimum
MAX	Maximum
NV	Normalverteilung

ABBILDUNGSVERZEICHNIS

FORMELVERZEICHNIS

TABELLENVERZEICHNIS

0. EINLEITUNG

Als Folge der zunehmenden Bedeutung von Qualifikationen und Wissen in den unterschiedlichen Lebensbereichen erlangt die Produktivität und Qualität des Bildungssystems eine gesellschaftliche Notwendigkeit (vgl. Klieme, Maag-Merki & Hartig, 2007, S. 5).

> „Gesellschaftlicher Wohlstand, soziale Kohäsion und Entwicklungschancen einer Gesellschaft hängen in entscheidendem Maße vom Bildungsstand ihrer Mitglieder ab" (Klieme & Leutner, 2006b, S. 876).

Der damit in Verbindung stehende Wunsch, von staatlicher Seite das Bildungssystem verstärkt zu steuern, macht es erforderlich, die Outputfaktoren empirisch zu erfassen (vgl. Klieme et al., 2007, S. 5 ff.). Dies wird nicht nur an den erworbenen Bildungsabschlüssen bemessen, sondern das summierte Produkt (bestehend aus Wissen, Fähigkeiten, Fertigkeiten, Einstellungen, Bereitschaften und Qualifikationen), das durch Bildungsprozesse entsteht, ist entscheidend. Mit dem zudem eingetretenen Bedeutungswandel von Bildung und Qualifikation geht es nicht mehr um das reine Erlangen und Weitergeben „eines festen Kanon[s] fachlicher Kenntnisse" (ebd., S. 9), sondern um eine kontextabhängige Wissensaneignung, die zum selbstständigen Lernen befähigt. Dementsprechend wird der Kompetenzerwerb der Schülerinnen und Schüler zum zentralen Ziel des Bildungssystems erklärt. Im Zuge dessen kann die Verbesserung des Bildungssystems durch die Kompetenzmessung erreicht werden, wie es das folgende Zitat auf den Punkt bringt:

> „Der Diagnostik von Kompetenzen kommt eine Schlüsselfunktion für die Optimierung von Bildungsprozessen und für die Weiterentwicklung des Bildungswesens zu" (Hartig & Jude, 2007, S. 17).

Allerdings gestaltet sich dieser Prozess theoretisch und methodisch schwierig (vgl. Klieme & Leutner, 2006b, S. 876 f.). Es besteht vor allem im Bereich der theoriegeleiteten Modellentwicklung zur Erfassung spezifischer Kompetenzen und der Entwicklung adäquater statistischer Modellierungsmethoden zur Umsetzung dieser Modelle ein Forschungsbedarf (vgl. Schweizer, 2006, S. 140 f.). Aufgrund der Erkenntnis, dass Bildungsprozesse prinzipiell in allen Lebensbe-

reichen stattfinden und aufeinander aufbauen, hat man es grundsätzlich mit dem Zusammenwirken mehrerer Faktoren bei der Untersuchung der Kompetenz zu tun. Um diesem Umstand gerecht zu werden, bedarf es eines breiten und im Idealfall längsschnittlich angelegten interdisziplinären und methodischen Zugangs, um die Ursache-Wirkungsbeziehung von Bildungsprozessen aufzuschlüsseln (vgl. Ditton & Reinders, 2011, S. 69 f.).

In Deutschland wird diesen umfassenden Forderungen erstmalig mit der Längsschnittstudie Nationales Bildungspanel (National Educational Panel Study, NEPS) nachgegangen, die sich auf der Organisationsstruktur anhand von sogenannten Etappen und Säulen mit der längsschnittlichen, interdisziplinären und methodischen Erfassung individueller Bildungsverläufe beschäftigt (vgl. Blossfeld, Schneider & Doll, 2009). Dazu werden in acht Etappen die Bildungsbiografien untersucht, um längsschnittliche Informationen zu den Bildungsverläufen und -übergängen zu gewinnen. Die fünf Säulen sind verantwortlich dafür, Instrumente zu entwickeln bzw. auszuwählen und zur Verfügung zu stellen (vgl. Blossfeld, von Maurice & Schneider, 2011a, S. 9 ff.).

Im Rahmen der NEPS-Studie werden Kompetenzen über die Schul- bzw. Lebenszeit erfasst. Dazu gehört die Entwicklung der Rechtschreibkompetenz im Verlauf der Sekundarstufe I (vgl. Frahm et al., 2011). Die Notwendigkeit einer solchen kompetenzorientierten Leistungsmessung der Rechtschreibung ergibt sich aus der Forschungslage. Die deutsche Rechtschreibung wird herkömmlich unter dem Paradigma der Dependenzthese als Abbild der gesprochenen Sprache verstanden. Als Idealfall gilt eine Eins-zu-Eins-Zuordnung von Phonem und Graphem. Alle davon abweichenden Schreibungen gelten als regelungsbedürftig und wurden erstmals im „Orthographischen Wörterbuch der deutschen Sprache“ von Konrad Duden 1902 festgelegt, das vom Grundsatz her die Basis für das heute gültige „amtliche Regelwerk“ liefert. Seit Kosog (1912) werden die Abweichungen von der Lauttreue als Ursache der Schwierigkeiten betrachtet, die das Erlernen der Rechtschreibung den Schülerinnen und Schülern bis heute bereiten. Dass Rechtschreibreformen das Lernen erleichtern, ließ sich wissenschaftlich bisher nicht nachweisen. Marx (1999) stellte in seinen Untersuchungen nach der jüngsten Reform im Jahr 1998 keine Verbesserung der Rechtschreibleistung fest. Ergebnisse aus jüngeren empirischen Untersuchungen zur Rechtschreibleistung deutscher Schülerinnen und Schüler weisen sogar auf eine zunehmende Verunsicherung hin (vgl. Blatt 2010, S. 101 f.). Im Zusammenhang damit wurde auf der Grundlage der Forschung zur Rechtschreibung ein sprachsystematisches Rechtschreibkompetenzmodell abgeleitet, das mithilfe eines sprachsystematischen Recht-

schreibtests (SRT) die differenzielle Erfassung der Rechtschreibkompetenz erlaubt. Die empirische Validierung des sprachsystematischen Rechtschreibkompetenzmodells erfolgte durch die Ergänzungsstudien Orthografie zu IGLU 2006 (IGLU-E) und zu HeLp 2007/08 (HeLp-E) für die Klassenstufen 4 und 5.

Auf diesen Forschungen aufbauend wurde das sprachsystematische Rechtschreibkompetenzmodell im Rahmen der NEPS-Studie für die Sekundarstufe I weiterentwickelt, um geeignete längsschnittliche Testinstrumente für die kompetenzorientierte Leistungsmessung zu entwickeln und adäquate Verfahren für die Modellierung der Rechtschreibkompetenz zu finden.

Die dieser Arbeit zugrundeliegenden Forschungsziele und -fragen werden aus Anlage, Zielsetzung und Auswertungsstand der etappenspezifischen Ergänzung der NEPS-Studie abgeleitet, um der Grundfrage nachzugehen, ob sich die Kompetenzstruktur der Schülerinnern und Schüler mit ihrer Kompetenzentwicklung verändert. Im Zuge der damit in Verbindung stehenden längsschnittlichen Erfassung der Rechtschreibkompetenz im Verlauf der Sekundarstufe I tauchen grundsätzliche Fragen auf, denen sich die vorliegende Arbeit als Forschungsziele widmet:

- Haben die in einer Entwicklungsstudie nachgewiesenen Gütekriterien für das Messinstrument SRT auch in der folgenden repräsentativen Haupterhebung Bestand?
- Welche Analyseverfahren eignen sich für die längsschnittliche Kompetenzentwicklung, um diese verlässlich und effizient zu erfassen und mögliche Änderungen der differenziellen Kompetenzstruktur aufzudecken?
- Welche Faktoren wirken sich auf die differenzielle Kompetenzentwicklung aus?

Aus den Forschungszielen ergeben sich vier konkrete Forschungsfragen, denen in der methodisch ausgerichteten Arbeit aus erziehungswissenschaftlicher Perspektive nachgegangen wird:

- *F1: Wie verhalten sich die in der Entwicklungsstudie und in der Haupterhebung in Klassenstufe 5 ermittelten Gütekriterien für den Rechtschreibtest zueinander?*
- *F2: Mit welchen methodischen Verfahren lässt sich die Entwicklung der Rechtschreibkompetenz verlässlich und effizient erfassen?*
- *F3: Verändert sich die Kompetenzstruktur von Klassenstufe 6 bis 7, und wenn ja, in welcher Weise?*

– *F4: Beeinflussen die Entwicklung in den Teilkompetenzen bzw. die Faktoren Geschlecht, Alter, sozioökonomischer Hintergrund, Migrationshintergrund und Schulform die Entwicklung der Rechtschreibkompetenz auf der Ebene des ganzen Wortes von Klassenstufe 6 und 7, und wenn ja, in welcher Weise?*

Zur Beantwortung der Forschungsfragen werden im Sinne der interdisziplinären Forschung empirische und didaktische Erkenntnisse für die quantitative Datenanalyse zur längsschnittlichen Erfassung der Rechtschreibkompetenz kombiniert.

Um die quantitative Datenanalyse zur längsschnittlichen Erfassung der Rechtschreibkompetenz in einen größeren Rahmen zu stellen, wird die empirische Bildungsforschung herangezogen und in Kapitel 1 überblicksartig vorgestellt. Weitergehend werden in Kapitel 2 die Leistungsmessung und Kompetenzmodellierung auf ihrer fachwissenschaftlichen Grundlage entwickelt, um daran anknüpfend in Kapitel 3 die dazugehörigen statistischen Analysemethoden zur längsschnittlichen Kompetenzerfassung darzustellen. Nach den allgemeinen Darlegungen zur Kompetenzerfassung und -modellierung wird der Fokus in Kapitel 4 auf die Rechtschreibung gerichtet und der Forschungsstand zur empirischen Erforschung von Rechtschreibleistung und -kompetenz aufgezeigt. Davon ausgehend wird das zur längsschnittlichen Erfassung der Rechtschreibkompetenz zentrale sprachsystematische Rechtschreibkompetenzmodell beschrieben und die empirische Validierung des Kompetenzmodells (Kapitel 5) dargestellt. Im anschließenden Kapitel 6 erfolgt die Vorstellung der NEPS-Studie, wobei auf die Rechtschreibtestung im Verlauf der Sekundarstufe I eingegangen wird. An die Darstellung und Diskussion des theoretischen, methodischen und empirischen Forschungsstandes knüpft die empirische Untersuchung an. Dabei wird einleitend das Forschungsvorhaben vorgestellt, wobei die Forschungsfragen und das dazugehörige methodische Vorgehen auf der Grundlage der in der NEPS-Studie gewonnenen Daten mit dem eingesetzten Testinstrument (Kapitel 7) benannt werden. Die gewonnenen Ergebnisse werden in Kapitel 8 unter mehreren Aspekten differenziert dargestellt und zu einer Gesamtaussage zusammengefasst sowie interpretiert. Abschließend werden die gewonnenen Ergebnisse in Kapitel 9 unter didaktischer und empirischer Perspektive diskutiert und darüber hinaus wird ein Fazit gezogen.

I. THEORETISCHER, METHODISCHER UND EMPIRISCHER FORSCHUNGSSTAND

Die längsschnittliche Erfassung der Rechtschreibkompetenz ist theoretischer, methodischer und empirischer Gegenstand dieser Arbeit und in der empirischen Bildungsforschung zu verankern. In dem insgesamt interdisziplinär angelegten Forschungsgebiet gilt die Erziehungswissenschaft als zentrale Bezugsdisziplin, weshalb unter einer erziehungswissenschaftlichen Perspektive der Fokus dieser Arbeit zunächst auf die empirische Bildungsforschung (Kapitel 1) gerichtet wird (vgl. Tippelt & Schmidt, 2010, S. 10).

Welche Modelle und Verfahren der empirischen Bildungsforschung zur längsschnittlichen Erfassung der Rechtschreibkompetenz zur Verfügung stehen, werden in den beiden nachfolgenden Kapiteln dargestellt: Dazu wird das Feld der Leistungsmessung und Kompetenzmodellierung (Kapitel 2) aufgegriffen. Inwiefern die längsschnittliche Kompetenzerfassung erfolgt, wird anhand der dazugehörigen statistischen Analyseverfahren (Kapitel 3) aufgezeigt.

1. EMPIRISCHE BILDUNGSFORSCHUNG

Der Zugang zur empirischen Bildungsforschung wird durch die Skizzierung der wichtigsten Stationen der historischen Entwicklung geschaffen (vgl. Hopf, 2004; Reinders, Ditton, Gräsel & Gniewosz, 2011b; Ditton, 2011), um anschließend die aktuellen Themen und Trends sowie die zentralen Vergleichsstudien dieses Forschungsgebiets zu betrachten.

1.1 HISTORISCHE ENTWICKLUNG UND AKTUELLES VERSTÄNDNIS

Die Entstehung und Entwicklung der empirischen Bildungsforschung wird ausgehend vom 20. Jahrhundert bis hin zur empirischen Wende im 21. Jahrhundert skizziert.

1.1.1 ENTWICKLUNG IM 20. JAHRHUNDERT

Zu Beginn des 20. Jahrhunderts ist die *experimentelle Pädagogik* (Meumann, 1914; Lay, 1918) in Deutschland prägend für die anfänglichen Versuche einer empirischen Bildungsforschung. Sie widmete sich mit einem psychologisch-naturwissenschaftlichen Verständnis Fragen der Bildung und stand damit der ganzheitlichen Auffassung von Bildung der geisteswissenschaftlich-philosophischen Erziehungswissenschaft entgegen. Das konkrete Bildungsverständnis der experimentellen Pädagogik lässt sich an deren Forschungslogik erkennen, bei der in experimentellen Laboruntersuchungen Unterrichtsformen simuliert wurden. Damit sollte sichergestellt werden, dass die Messung des reinen Lernprozesses ohne störende Einflüsse erfolgt. Erkenntnisse zu Unterrichtsformen und Bildungsprozessen, die aus diesen Untersuchungsergebnissen resultieren, sollten im Sinne der Forschungslogik in die Schulpraxis übertragen werden. Allerdings scheiterte der Transfer experimenteller Erkenntnisse in die schulische Praxis, da das aus Experimenten gewonnene empirische Wissen einer anderen Logik unterliegt, als sie in der Bildungspraxis vorherrscht. Dies hatte zur Konsequenz, dass sich die Forschung von den künstlich erzeugten Laboruntersuchungen abwandte und dazu überging, das natürliche Bildungsumfeld zu evaluieren.

Weitergehend konkretisierte sich in den 1960er-Jahren ein erster Vorläufer der empirischen Bildungsforschung, und zwar infolge der intensiven Diskussion über die zukünftige Gestaltung des Bildungssystems, die am Ende des Zweiten Weltkrieges begann und seither andauerte. Die vorrangig ökonomisch orientierte Diskussion über das Bildungssystem sah das Kapital und die Bildung als entscheidenden Faktor für die wirtschaftliche Entwicklung Deutschlands an (Edding, 1963). Picht (1964) machte zudem auf die Notwendigkeit einer international vergleichenden Bildung aufmerksam. Nach Dahrendorf (1965) bestand im Sinne einer aktiven Bildungspolitik Handlungsbedarf mit dem Ziel, den Zugang zur Bildung zu erleichtern und jedem Menschen das Recht zu geben, sich zu bilden und somit an der Gesellschaft teilzuhaben. In diesem Zusammenhang wurde 1965 der *Deutsche Bildungsrat* gegründet, dessen durchgeführte Studien und erstellte Gutachten zu den Leitfragen der Bildungsforschung bis heute Gültigkeit besitzen. Ein zentrales Anliegen stellte dabei die Forderung nach einer Steigerung der Bildungsbemühungen und der Berücksichtigung von Chancengleichheit im Bildungssystem dar. Der Blick richtete sich jetzt auch auf die Bildungsvoraussetzungen, die Menschen vor ihrem Eintritt in die institutionalisierte Bildung mitbringen und die von der sozialen Herkunft abhängig sind (Coleman, 1966). Darüber hinaus sind die strukturellen und inhaltlichen Merkmale

der Bildungsinstitution zur Steuerung und Steigerung der Lernleistung der Schülerinnen und Schüler in die Untersuchungen einzubeziehen (Flechsig, Tütken, Riedel, Thiersch & Skowronek, 1968). Mit dem veränderten Verständnis von Bildung war die Annahme verbunden, dass Bildungsprozesse nicht durch angeborene Merkmale wie Intelligenz determiniert sind, sondern durch Lernmöglichkeiten gefördert und weiterentwickelt werden können (Roth & Aebli, 1969). Zusammenfassend hat sich eine neue Sichtweise auf Bildung ergeben, die in das Bildungswesen integriert werden sollte. Durch diesen Wandel des Bildungsbegriffs wurden Lernumwelten zu einem zentralen Forschungsanliegen, um durch geeignete Bildungsangebote inner- und außerhalb von Bildungsinstitutionen Bildungsprozesse zu ermöglichen.

Aus den bisher dargestellten Entwicklungen heraus etablierte sich eine empirische Bildungsforschung. Wie der Deutsche Bildungsrat eine solche Forschung in den 1970er-Jahren verstand, ist im folgenden Zitat wiedergegeben:

> „Man kann Bildungsforschung in einem weiteren und engeren Sinne auslegen. Im engeren Sinne hat es sie als Unterrichtsforschung schon immer gegeben. Im weiteren Sinne kann sie sich auf das gesamte Bildungswesen und seine Reform im Kontext von Staat und Gesellschafft beziehen einschließlich der außerschulischen Bildungsprozesse. Wie weit oder eng aber auch die Grenzen der Bildungsforschung gezogen werden, es sollte nur dann von Bildungsforschung gesprochen werden, wenn die zu lösende Aufgabe, die Gegenstand der Forschung ist, theoretisch oder empirisch auf Bildungsprozesse (Lehr-, Lern-, Sozialisations- und Erziehungsprozesse), deren organisatorische und ökonomische Voraussetzungen oder Reform bezogen ist“ (Deutscher Bildungsrat, 1974, S. 23).

Die empirische Bildungsforschung kann also nach Auffassung des Deutschen Bildungsrats in einem weiteren und engeren Sinn verstanden werden. Im weiteren Sinne ist das Gesamtbild des Bildungswesens gemeint, wozu inner- und außerschulische Institutionen sowie die gesellschaftlich und bildungspolitisch bedingten Maßnahmen in Form von Schulsystemen und deren Reformen gehören. Dagegen erfolgt im engeren Sinne die Auseinandersetzung mit dem Unterricht und den damit verbundenen Bildungsprozessen. Darüber hinaus kann nur von Bildungsforschung ausgegangen werden, wenn der zu untersuchende Gegenstand sich auf Bildungsprozesse bezieht.

Darüber hinaus vertraten seit den 1970er-Jahren internationale und nationale Bildungsexperten die Auffassung, dass in der modernen Gesellschaft Kernkompetenzen benötigt werden, die objektiv gemessen werden können. Aus diesem Grund fokussierte sich das Interesse des Deutschen Bildungsrates auf die Entwicklung geeigneter Verfahren zur Leistungsdiagnostik und Leistungsrückmeldung auf der Grundlage inhaltlich definierter Leistungsstandards, die sich auch zur objektiven Beurteilung für die Prognose des Bildungsverlaufs und der Übergangsempfehlung eignen sollten (Fricke, 1974).

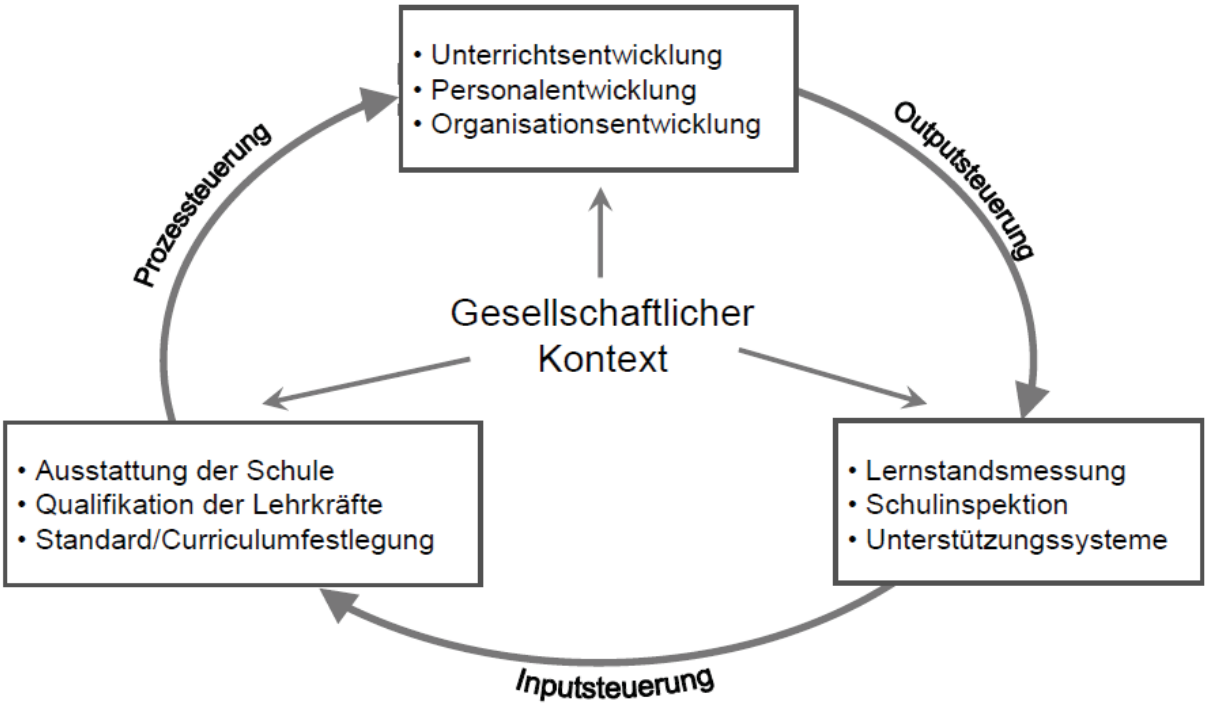
1.1.2 EMPIRISCHE WENDE IM 21. JAHRHUNDERT

Der Aufschwung der empirischen Bildungsforschung seit 2000 ging sowohl mit theoretischen als auch methodischen Entwicklungen einher. So stützt sich die Bildungspolitik seitdem auf empirisch gewonnene Daten aus Schülerleistungsmessungen, um das Bildungssystem zielgerichtet weiterzuentwickeln und zu gestalten. Daher spricht man auch von einer empirischen Wende in der Bildungsforschung und im Bildungssystem. Das gestiegene Interesse an der Bildungsforschung zeigt sich daran, dass die Ergebnisse von repräsentativen empirischen Studien auf der politischen und öffentlichen Ebene diskutiert werden. Durch die öffentliche Wahrnehmung erlangt die Bildungsforschung einen neuen Stellenwert und spricht dem Bildungssystem zugleich eine gesellschaftliche Bedeutung zu, wobei sie gleichermaßen für den Erfolg und Misserfolg herangezogen wird (vgl. Zedler & Döbert, 2010, S. 34).

Eine zentrale Rolle für die weitere Entwicklung des Verständnisses der empirischen Bildungsforschung in den letzten Jahren spielte unter anderem der sogenannte „PISA-Schock“ (vgl. Reinders, Ditton, Gräsel & Gniewosz, 2011a, S. 9), der eine intensive Diskussion über die Leistungsfähigkeit des Schulsystems auslöste (vgl. Gräsel, 2011, S. 16). Die durch den Schock veranlassten bundesweiten Maßnahmen auf politischer und wissenschaftlicher Ebene zur Verbesserung des Bildungssystems ergänzen oder ersetzen die bis dahin richtungsweisende „Input-Orientierung“ durch eine „Output-Orientierung“ (vgl. Hornberg & Bos, 2007, S. 178 ff.). Mit der eingetretenen Wende, den Fokus auf den Output des Bildungssystems zu legen, sind die individuellen Merkmale der Schülerinnen und Schüler, die zum lebenslangen Lernen und zur gesellschaftlichen Beteiligung beitragen, besonders relevant geworden. Der Schule wird über den Kompetenzerwerb hinaus die Verantwortung dafür zugesprochen, die Schülerinnen und Schüler zu einer kulturellen Teilhabe an der Gesellschaft zu befähigen und den sozialen

Zusammenhalt der Gesellschaft zu gewährleisten. Dazu wird besonders auf das erfolgreiche Zustandekommen von Bildungsprozessen geachtet. Es besteht eine staatliche Verpflichtung, mithilfe von vielseitigen Indikatoren zu kontrollieren, ob diese Ziele tatsächlich erreicht werden und somit die Qualität des Bildungssystems gesichert ist. Aufgrund dieser Verpflichtung erfolgt unter Zuhilfenahme der empirischen Bildungsforschung ein regelmäßiges und systematisches Systemmonitoring zur Erfassung der Bedingungen und Ergebnisse des Bildungssystems (vgl. Hornberg & Bos, 2007, S. 155). Auf der Grundlage des Systemmonitorings besteht für den Staat die Möglichkeit, das Bildungswesen auf der Input-, Prozess- und Outputebene zu steuern, wie es in Abbildung 1.1 veranschaulicht ist.

Abbildung 1.1: Kreislauf der Steuerung im Bildungswesen



(Bos, Holtappels & Rösner, 2006, S. 83)

Der Kreislauf zur Steuerung des Bildungswesens wird durch den gesellschaftlichen Kontext auf der Input-, Prozess- und Outputebene beeinflusst, also beispielsweise durch gesellschaftliche Probleme („PISA-Schock") oder aufgrund der sozialen Merkmale der Schülerinnen und Schüler. Auf der Inputebene kann der Staat durch die Bereitstellung ausreichender materieller

und personeller Ressourcen das Bildungswesen steuern, wozu die Ausstattung der Schulen sowie die Qualifikation der Lehrkräfte gehören. Ebenso liefert die Inputebene eine Orientierung für die Bildungsprozesse, indem durch curriculare Vorgaben festgelegt wird, wie und was die Schülerinnen und Schüler während ihrer Schulzeit lernen sollen. Neben der Bereitstellung der notwendigen Ressourcen hat sich gezeigt, dass auch die Prozessebene zur Erreichung der Bildungsziele gesteuert werden muss (vgl. Rolff, 1993). Dementsprechend wird eine Unterrichts-, Personal und Organisationsentwicklung zur Steuerung der Prozessebene benötigt, um anknüpfend an die Inputfaktoren die bestmöglichen Bedingungen für die Bildungsprozesse zu erreichen. Inwiefern sich die Steuerungsbemühungen auf der Input- und Prozessebene auswirken, lässt sich auf der Outputebene überprüfen, indem ergänzend Lernstandsmessungen, Schulinspektionen sowie Unterstützungssysteme zur Outputsteuerung des Bildungssystems in Anspruch genommen werden (vgl. Hornberg & Bos, 2007, S. 178 ff.).

Die empirische Bildungsforschung bezieht dafür unterschiedliche Disziplinen ein und setzt auf eine interdisziplinäre Zusammenarbeit (vgl. Reinders et al., 2011a, S. 10). Die empirische Bildungsforschung ist jedoch nicht eindeutig auf bestimmte Forschungsinhalte festgelegt (vgl. Tippelt & Schmidt, 2010, S. 9; Reinders et al., 2011a, S. 9; Zedler & Döbert, 2010, S. 23). Ein Blick in die Literatur zur empirischen Bildungsforschung offenbart schnell die Vielfältigkeit der Forschungsthemen (vgl. Reinders et al., 2011b, S. 10). Insgesamt zeigt sich jedoch, dass die empirische Bildungsforschung ein international orientiertes Interesse bezüglich der Fragen zur Bildung und Bildungsqualität verfolgt (vgl. Reinders et al., 2011a, S. 9).

1.1.3 AKTUELLES VERSTÄNDNIS

Versucht man die aktuelle Auffassung der empirischen Bildungsforschung zu konkretisieren, stellt sich dies als ein schwieriges Unterfangen dar, da sie vom zugrundeliegenden Verständnis der jeweiligen Disziplin abhängig ist. Dem heutigen Stand entspricht vor allem das Bildungsverständnis aus den Sozialwissenschaften, wie es im folgenden Zitat zum Ausdruck kommt:

„Das Ziel empirischer Sozialforschung ist es, Kenntnisse über die soziale Realität zu erlangen. Empirische Bildungsforschung begrenzt den Untersuchungsgegenstand auf Lern- und Bildungsprozesse als Ausschnitt dieser sozialen Realität [...]. Es sollen mit Mitteln der Beobachtung dieser Bildungsrealität Kenntnisse gewonnen werden, wie Bildung funktioniert und unter welchen Bedingungen dieser Prozess (optimal) verläuft" (Reinders & Ditton, 2011, S. 45).

Die empirische Bildungsforschung stellt einen Teilbereich der Sozialwissenschaften dar und untersucht die soziale Realität in Hinblick auf Lern- und Bildungsprozesse. Mithilfe der quantitativen und qualitativen Beobachtung der Bildungsrealität werden Kenntnisse über die Funktion von Bildung und die Entwicklung von Bildungsprozessen gewonnen. In Bezug auf die Orte, an denen Bildungsprozesse stattfinden, wird zwischen *formeller, nicht-formeller* und *informeller Bildung* unterschieden. Darunter ist nach Harring, Rohlfs & Palentien (2007) Folgendes zu verstehen:

- Mit *formeller Bildung* sind die intendierten Lernprozesse gemeint, die in den dafür vorgesehenen formalen Institutionen des Bildungssystems (Schule, Ausbildung, Hochschule) erfolgen. Durch einen reglementierten und zielgerichteten Ablauf in diesen Institutionen führt das formale Lernen zum Erwerb von Qualifikationen, die im Weiteren die soziale und berufliche Entwicklung der Schülerinnen und Schüler mitbestimmen.
- Die *nicht-formelle Bildung* zeichnet sich durch eine freiwillige Entscheidung über die Inanspruchnahme von Bildungsangeboten aus. Hierbei steht der Erwerb von sozialen und personalen Kompetenzen im Fokus, die zur politischen und gesellschaftlichen Teilhabe beitragen.
- *Informelle Bildung* meint situatives Lernen, also indirekte, ungeplante und beiläufige Lernprozesse, die beispielsweise in Rahmen der Familie oder Peer Groups erfolgen. Mit der informellen Bildung wird die Grundlage für formelle und nicht-formelle Bildungsprozesse geschaffen.

Gräsel (2011) konkretisiert das Verständnis der empirischen Bildungsforschung, indem sie einen Bezug zur formellen bzw. institutionalisierten Bildung herstellt.

„Die Empirische Bildungsforschung untersucht die Bildungsrealität in einer Gesellschafft, wobei der Schwerpunkt auf der institutionalisierten Bildung liegt. Bildungsforschung fragt im Kern, wie Bildungsprozesse verlaufen, wer welche Qualifikationen und Kompetenzen im Bildungssystem erwirbt, wovon dieser Qualifikations- und Kompetenzerwerb abhängig ist, und welche Auswirkungen er hat“ (Gräsel, 2011, S.12).

Demnach geht es um in konkreten Lernumgebungen beobachtbare Bildungsprozesse, die vorrangig in dafür vorgesehenen Institutionen ablaufen. Im Fokus der empirischen Bildungsforschung steht, welche Kompetenzen durch die Bildung erlangt werden, wie sie sich entwickeln und wovon dies abhängt.

Nach Gräsel (2011) haben sich aufgrund der Definition der Bildungsforschung durch den Deutschen Bildungsrat bis heute drei zentrale Merkmale etabliert:

1. Problemorientierung
2. Interdisziplinarität
3. Verwendung empirischer Forschungsmethoden

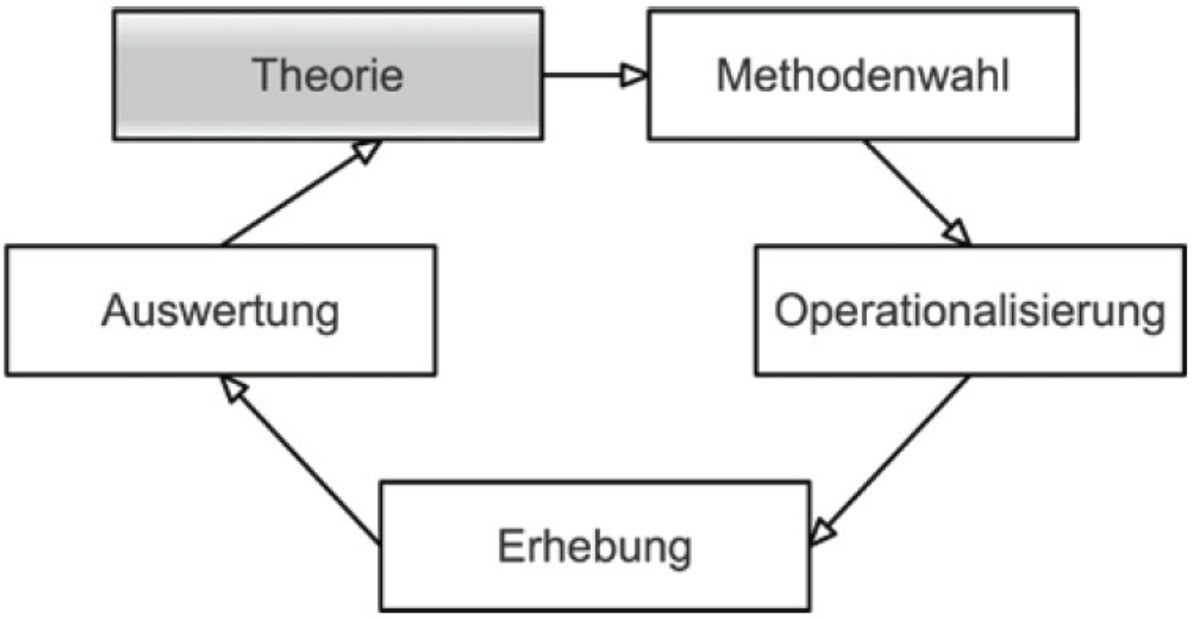
Die empirische Bildungsforschung liefert wissenschaftliche Erkenntnisse für die Analyse von Problemen im Bildungssystem mit dem Ziel, eine Verbesserung zu bewirken. Ebenso werden rationale Begründungen für bildungspraktische und -politische Entscheidungen zur Verfügung gestellt, um die Bedingungen auf der Ebene des Bildungssystems, einzelner Institutionen sowie des individuellen Lernens zu verbessern (vgl. Tippelt & Schmidt, 2010, S. 9). Für eine solche Forschung ist der Einbezug verschiedener Disziplinen mit ihren theoretischen und methodischen Zugängen notwendig, um mittels der Interdisziplinarität Antworten auf die komplexen Forschungsfragen der empirischen Bildungsforschung zu finden. Dies wird nach dem aktuellen Verständnis von empirischer Bildungsforschung dadurch erreicht, dass empirische Forschungsmethoden verwendet werden, die sich an den Standards der Sozialwissenschaften orientieren und quantitative sowie qualitative Vorgehensweisen beinhalten. Inwiefern sich die quantitativen und qualitativen Forschungsmethoden der empirischen Bildungsforschung voneinander abgrenzen, wird in dem folgenden Zitat verdeutlicht:

„Quantitative Methoden folgen dem Prinzip der Deduktion und sind theorie- oder hypothesenprüfend angelegt. Qualitative Methoden fokussieren den Einzelfall und leiten bei entsprechendem Interesse eine Theorie oder ein Konzept aus der Einzelfallbetrachtung ab“ (Reinders & Ditton, 2011, S. 48).

Die quantitativen Forschungsmethoden dienen der Theorie- und Hypothesenprüfung (Deduktion) und die qualitativen Forschungsmethoden der Theoriegenerierung (Induktion). Die deduktiven Methoden überprüfen demnach eine allgemeine Theorie bzw. Hypothese an besonderen Fällen und in der Umkehrung leiten die induktiven Methoden aus besonderen Fällen eine allgemeine Theorie bzw. Hypothese ab (vgl. Reinders & Ditton, 2011, S. 47 f.).

Der unterschiedliche Forschungsablauf der quantitativen und qualitativen Forschungsmethoden der empirischen Bildungsforschung wird in der folgenden Abbildung schematisch dargestellt.

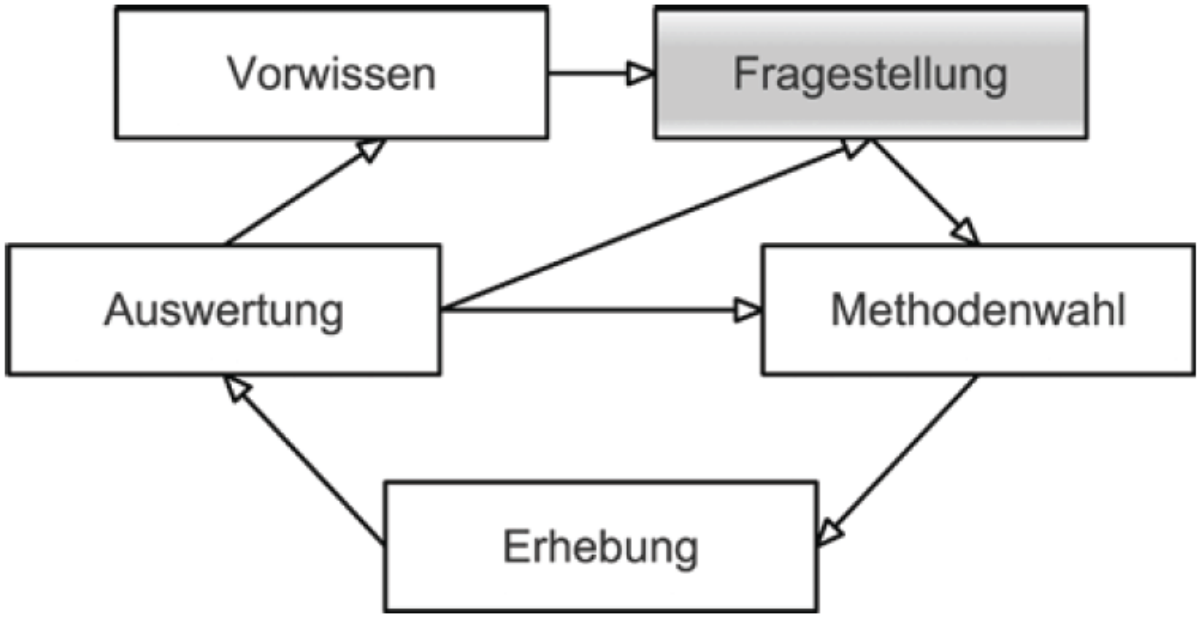
Abbildung 1.2: Schematischer Ablauf quantitativer Bildungsforschung



(Reinders & Ditton, 2011, S. 49)

Bei der quantitativen Forschung steht die Prüfung von Theorien bzw. Hypothesen im Fokus, an der sich die Methodenwahl zur Beobachtung der mit der Theorie bzw. Hypothese definierten Merkmale orientiert. Diese Merkmale werden durch Operationalisierung, beispielsweise mithilfe von Fragen oder Testaufgaben, messbar. Die Auswertung der erhobenen Daten liefert Ergebnisse, um die zugrundeliegende Theorie bzw. Hypothese zu validieren bzw. zu falsifizieren (vgl. Reinders & Ditton, 2011, S. 48 f.).

Abbildung 1.3: Schematischer Ablauf qualitativer Bildungsforschung



(Reinders & Ditton, 2011, S. 50)

Auf der Grundlage von Vorwissen ergeben sich bei der qualitativen Forschung Fragestellungen, die im Verlauf des Forschungsprozesses explorativ beantwortet werden. Dies bedeutet, dass der Forschungsablauf vorab nicht exakt festgelegt wird, sondern dass sich Fragestellung und Methodenauswahl im Untersuchungsverlauf abhängig von den laufenden Zwischenergebnissen weiterentwickeln. So beinhaltet die qualitative Forschung einen dynamischen Forschungsablauf, der zur explorativen Beantwortung von auf Vorwissen basierenden Forschungsfragen dient.

Ebenso ist die Kombination der quantitativen und qualitativen Forschungsmethoden im Sinne einer Triangulation möglich (vgl. Gläser-Zikuda, 2011, S. 118). Mit einer so ausgerichteten em-

pirischen Bildungsforschung können der Output des Bildungssystems und die darauf einwirkenden Einflussfaktoren unter Zuhilfenahme von Mess- und Befragungsverfahren erfasst werden.

1.1.4 AKTUELLE TRENDS

Aktuell lassen sich Trends in der empirischen Bildungsforschung ausmachen, die sich zusammenfassend mit (1) internationalen und nationalen Leistungsvergleichen, (2) sozialer Selektivität des Bildungswesens, (3) Bildungspanel – Längsschnittstudien, (4) Unterrichtsqualität – Förderung von Kompetenzen sowie (5) Weiterentwicklung der Forschungsmethoden beschäftigen (vgl. Gräsel, 2011, S. 15 ff.).

Anhand der umfassenden Querschnittsuntersuchungen der OECD (Organisation for Economic Co-operation and Development) ist es möglich, den Status quo des Bildungssystems darzustellen, um mittels internationaler und nationaler Leistungsvergleiche (1) empirisch geleitete Diskussionen über die Qualität von Bildungsprozessen zu führen. Für den Vergleich der Leistungen werden Bildungsstandards (Kapitel 2.2) herangezogen. Zudem sind aufgrund der Leistungsergebnisse Rankings möglich, die Hinweise auf Disparitäten im Bildungssystem geben. Darunter fallen vor allem die soziale Selektivität (2) und die Chancenungleichheit des Bildungssystems (Allmendinger, Ebner & Nikolai, 2010), die vielfach empirisch nachgewiesen wurden (Baumert et al., 2001; Bos et al., 2003). Demnach sind der Bildungsabschluss und der Kompetenzerwerb der Schülerinnen und Schüler von ihrer sozialen Herkunft abhängig. Dem Bildungssystem gelingt es nicht, die ungleichen sozialen Voraussetzungen auszugleichen, wodurch Gruppen mit bestimmten sozioökonomischen Merkmalen nur eingeschränkte Aufstiegsmöglichkeiten erhalten. Der dritte Trend ist dadurch bedingt, dass mit Ergebnissen aus Querschnittuntersuchungen keine Aussagen über Bildungsprozesse und -verläufe zu treffen sind. Für diesen Zweck sind Bildungspanel bzw. Längsschnittstudien (3) erforderlich, z. B. das Nationale Bildungspanel in Deutschland (Kapitel 6.1) (Blossfeld et al., 2009). In den Längsschnittstudien werden nicht nur Kompetenzen erfasst, sondern auch schulische und außerschulische Einflussfaktoren für die Kompetenzentwicklung. Ein grundlegender Aspekt ist dabei die Erforschung der Unterrichtsqualität (4) als ein zentraler Einflussfaktor auf die Kompetenzentwicklung. Die Unterrichtsqualität wird im Rahmen der „Lehr-Lern-Forschung" bzw. Unterrichtsforschung betrachtet. Diese Forschung zielt auf die Verbesserung der Unterrichtsqua-

lität, indem die Lerninhalte und -methoden interdisziplinär hinterfragt und in praxisrelevante Unterrichtskonzeptionen überführt werden. Darüber hinaus liefern Metaanalysen (Hattie, 2009) Befunde, die zu den Bedingungen erfolgreichen Unterrichts Auskunft geben (vgl. Zusammenfassung bei Köller, 2012). Die Unterrichtsqualität hängt demnach in zentraler Weise von der Professionalität des Lehrers ab, die wiederum in der Lehrerforschung untersucht wird. Ein weiteres Anliegen der empirischen Bildungsforschung ist die Weiterentwicklung der Forschungsmethoden (5), d. h. der Erhebungs- und Analyseverfahren, um einen hohen forschungsmethodischen Standard zu gewährleisten. Zentral ist hierbei eine verlässliche Kompetenzmessung. Sie wirft, insbesondere auch in Hinblick auf eine längsschnittliche Erfassung, viele methodische Fragen auf (vgl. Gräsel, 2011, S.15 ff.).

1.2 QUER- UND LÄNGSSCHNITTLICHE VERGLEICHSSTUDIEN

Im Folgenden werden die quer- und längsschnittlichen Designs der Vergleichsstudien in der empirischen Bildungsforschung vorgestellt. Dazu wird insbesondere das Design der Längsschnittstudien von weiteren Studiendesigns abgegrenzt, um die spezifischen Möglichkeiten zur längsschnittlichen Erfassung von Kompetenzen zu verdeutlichen. Neben der strukturellen Beschreibung von Längsschnittstudien erfolgt die Entwicklung einer inhaltlichen Charakterisierung von Vergleichsstudien, um abschließend Längsschnittuntersuchungen zur Rechtschreibkompetenz anhand von Kategorien zu beschreiben und deren methodisches Vorgehen herauszustellen.

Zur Erfassung der Ergebnisse von Bildungsprozessen existieren zwei Arten von Vergleichsstudien bzw. Large-Scale-Assessments. Die erste Art der Vergleichsstudien orientiert sich an den Lehrplänen oder Bildungsstandards der jeweiligen Fächer in bestimmten Jahrgangsstufen, um die Ergebnisse von Bildungsprozessen zu erfassen. Zu dieser Studienart zählen beispielsweise Studien wie TIMSS (Trends in International Mathematics and Science Study) oder länderbezogene Vergleichsarbeiten wie VERA (*Ver*gleichs*ar*beiten in der 3. und 8. Jahrgangsstufe). Die zweite Art der Vergleichsstudien erfasst zu einem bestimmten Zeitpunkt die erworbenen Kompetenzen der Schülerinnen und Schüler in bestimmten Fächern als Ergebnis von Bildungsprozessen, die zum lebenslangen Lernen und zur gesellschaftlichen Teilhabe beitragen. Zu dieser Studienart zählen z. B. die Literacy-Studien PIRLS/IGLU (Progress in International Reading Literacy/Internationale Grundschul-Lese-Untersuchung) oder PISA (Programme

for International Student Assessment), die mit anwendungsbezogenen und realitätsnahen Tests prüfen, inwiefern die Schülerinnen und Schüler in der Lage sind, durch ihr Wissen, ihre Fertigkeiten und Strategien die in den Testaufgaben gestellten Anforderungen zu erfüllen (vgl. Drechsel, Prenzel & Seidel, 2009, S. 363 ff.). Diese beiden Arten der Vergleichsstudien basieren auf einem Grundmodell zur Darstellung der Wirkungsweise von Bildungssystemen, das die Input-, Prozess- und Outputebene umfasst (vgl. Dunkin & Biddle, 1974). Vor allem die Outputebene zählt zu den Interessensgebieten der empirischen Bildungsforschung, da hierdurch Erkenntnisse über die Schülerleistungen sowie die Erträge des Bildungssystems ermöglicht werden (vgl. Drechsel et al., 2009, S. 355 f.). Das Erfassen von Leistungsständen in den Vergleichsstudien „kann entsprechend als kompetenzorientierte Leistungsmessung bezeichnet werden“ (Voss, 2009, S. 14), da sie auf empirisch validierten Kompetenzmodellen basiert.

1.2.1 STUDIENDESIGNS

Bei den Vergleichsstudien wird zwischen den Studiendesigns der Quer- und Längsschnittuntersuchungen unterschieden (vgl. Häder, 2010, S. 115 ff.; Diekmann, 2007, S. 304). Im Wesentlichen unterscheiden sich die Designs durch die Veränderungen der Stichprobenzusammensetzung im Erhebungszeitraum und der Anzahl der Erhebungen, wobei auch Mischformen durch Kopplung von Quer- und Längsschnittuntersuchungen möglich sind. Die beiden Studiendesigns sind in Tabelle 1.1 gegenübergestellt:

Tabelle 1.1: Untersuchungsdesigns in empirischen Bildungsstudien

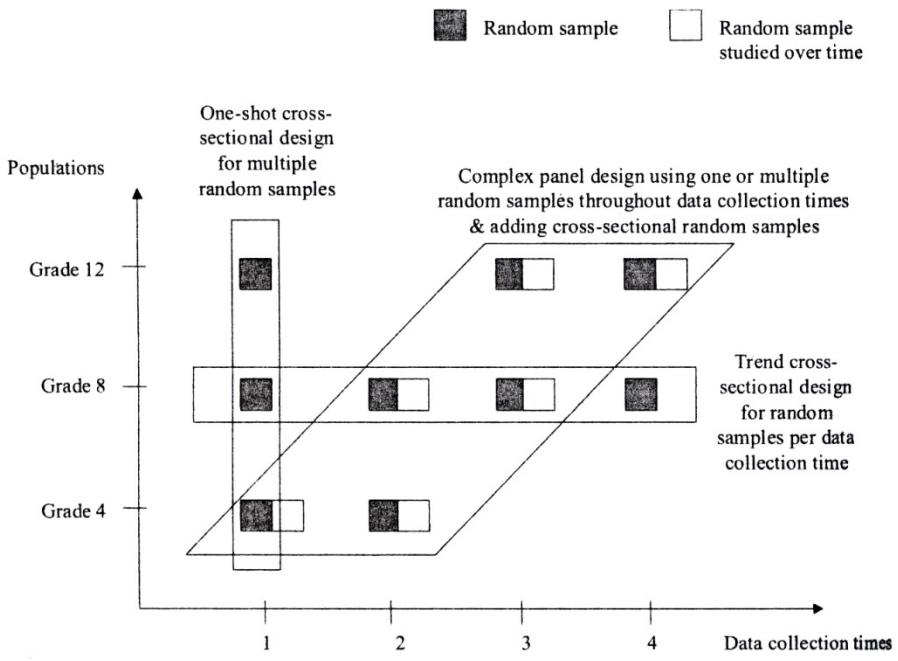
Design	Datenerhebungen	Zielgruppen	Stichproben
Querschnittlich			
Eine Erhebung	1	≥1 (z. B. Klasse 4, 8, 12)	Zufallsstichproben für jede Population zu jedem Erhebungszeitpunkt
In mehrjährigen Runden wiederkehrende Erhebungen (Trend)	>1	≥1	Zufallsstichproben für jede Population zu jedem Erhebungszeitpunkt
Längsschnittlich			
Einfaches Panel	>1	1	Eine Zufallsstichprobe wird zu Beginn gezogen und zu allen Erhebungszeitpunkten erneut getestet/befragt
Komplexes Panel	>1	≥1	Mehrere Zufallsstichproben werden zu Beginn gezogen und zu allen Erhebungszeitpunkten erneut getestet/befragt

(Drechsel et al., 2009, S. 364)

Das Design der Querschnittuntersuchungen ist dadurch gekennzeichnet, dass es sich auf einen Erhebungszeitpunkt oder eine jährlich wiederkehrende Erhebung (Trend) bezieht. Bei dem querschnittlichen Design stehen eine oder mehrere Zielgruppen im Fokus, wobei zu jedem Messzeitpunkt eine neue Stichprobe untersucht wird und sich so Veränderungen auf der Aggregatebene (z. B. Klassen-, Schul- oder Landesebene) nachweisen lassen. Handelt es sich dagegen um wiederholte Erhebungen bei einer identischen Stichprobe, spricht man von Längsschnitt- bzw. Paneluntersuchungen, bei denen sich Veränderungen auf der individuellen Ebene aufzeigen lassen. Einfache Panelstudien konzentrieren sich dabei auf eine Zielgruppe, wohingegen komplexe Panelstudien die mehrfache Erhebung unterschiedlicher Zielgruppen ermöglichen. Die durch Tests gewonnenen Daten werden mithilfe von Fragebögen, Interviews oder Beobachtungen durch Kontext- bzw. Hintergrundinformationen (z. B. soziale Merkmale, motivationale Aspekte) ergänzt. Ziel der Vergleichsstudien ist es, anhand von Analysen eine Beziehung zwischen den Leistungsdaten und den erklärenden Hintergrundinformationen herzustellen, um so Kompetenzunterschiede aufgrund von unterschiedlichen Lern- und Entwicklungsbedingungen zu untersuchen (vgl. Drechsel et al., 2009, S. 363 ff.).

Das Zusammenspiel der beiden Designs wird in der nachfolgenden Abbildung 1.4 anhand eines Beispiels verdeutlicht.

Abbildung 1.4: Studiendesigns im Vergleich



(Seidel & Prenzel, 2008, S. 282)

Für eine einmalige Erhebung und Beschreibung der Ergebnisse von Bildungsprozessen in den Jahrgangsstufen 4, 8, und 12 (Zielgruppen) mit zufälligen Stichproben eignet sich das querschnittliche Design (One-shot cross-sectional design). Soll dagegen z. B. die Beschreibung der Bildungsergebnisse in der Jahrgangsstufe 8 mehrfach vorgenommen werden, bieten sich wiederkehrende Erhebungen unterschiedlicher Stichproben in Form eines Trends (Trend cross-sectional design) an. Ist dagegen z. B. die längsschnittliche Beschreibung der Ergebnisse von Bildungsprozessen mehrerer Stichproben das Ziel, eignet sich ein komplexes Paneldesign (Complex panel design), wenn eine Entwicklung über die drei Jahrgangstufen oder innerhalb einer Jahrgangsstufe aufgezeigt werden soll.

Mit den Designs der Querschnitt- und Längsschnittuntersuchungen sind mehrere Einschränkungen verbunden. Mit Festlegung des Studiendesigns werden gleichzeitig die Interpretati-

onsmöglichkeiten der auf diesem Weg gewonnen Daten festgelegt. So ermöglichen Querschnittuntersuchungen lediglich die deskriptive Beschreibung der erfassten Ergebnisse von Bildungsprozessen. Hingegen erlaubt die Längsschnittuntersuchung die Herstellung kausaler Beziehungen zwischen den Leistungsdaten und den Hintergrundinformationen (vgl. Drechsel et al., 2009, S. 363). Zudem bieten Längsschnittuntersuchungen die Möglichkeit, Lernentwicklungen abzubilden. Sie sind jedoch kosten- und zeitintensiver als Querschnittstudien (vgl. Bos & Gröhlich, 2009, S. 7). Ebenso kommt es infolge der mehrmaligen Erhebungen meistens zu systematischen Ausfällen in der Stichprobe bzw. zu einer Panelmortalität (vgl. Diekmann, 2007, S. 309). Demnach fallen einzelne Gruppen oder Personen aus unterschiedlichen Gründen weg und müssen teilweise durch andere Personen ersetzt werden. Gleichzeitig kann es aufgrund der wiederholten Messungen zu einem Paneleffekt kommen, womit die Wiedererkennung und Übung von Testaufgaben gemeint ist. Mittels eines alternierenden Panels kann dieses Problem kompensiert werden, da die Erhebungen bei einer ausreichend großen Stichprobe an zwei Gruppen (G1 und G2) durchgeführt wird.

Abbildung 1.5: Alternierendes Panel

	t_1	t_2	t_3	t_4	t_5
G1	X		X		X
G2		X		X	

(Schnell, Hill & Esser, 2011, S. 235)

Die Verwendung einer identischen (repräsentativen) Stichprobe erlaubt zwar Aussagen auf der individuellen Ebene, aber berücksichtigt dabei keine eventuellen Veränderungen in der Grundgesamtheit. Dieses Problem kann durch ein rotierendes Panel umgangen werden, weil zu jedem Zeitpunkt eine neue (repräsentative) Stichprobe aus der Grundgesamtheit gezogen wird und die früheste Gruppe aus dem Panel ausscheidet. Ebenso ist die Erhebung zeitlich begrenzt und endet mit der letzten Ursprungsgruppe.

Abbildung 1.6: Rotierendes Panel

	t_1	t_2	t_3	t_4	t_5
G1	X	X	X	X	X
G2	X	X	X	X	
G3	X	X	X		
G4	X	X			
G5	X				
G6		X	X	X	X
G7			X	X	X
G8				X	X
G9					X

(Schnell et al., 2011, S. 236)

Um die Vorteile beider Designs nutzen zu können, wird ein geteiltes Panel eingesetzt, welches sowohl eine Quer- als auch eine Längsschnittuntersuchung umfasst und dadurch den Panel-effekt reduziert. Die besonderen Panelformen führen jedoch zu einem hohen forschungsorganisatorischen Aufwand (vgl. Schnell et al., 2011, S. 233 ff.; Häder, 2010, S. 119 ff.)

Abbildung 1.7: Geteiltes Panel

	t_1	t_2	t_3	t_4	t_5
G1	X	X	X	X	X
Q1	X				
Q2		X			
Q3			X		
Q4				X	
Q5					X

(Schnell et al., 2011, S. 236)

Gleichzeitig ist die Wahl des Erhebungsdesigns eine Frage des gewünschten Datenmaterials, da es zur Sammlung aussagekräftiger Daten zur Prüfung von Hypothesen oder Beschreibung sozialer und ökonomischer Verhältnisse dient. Hierzu werden Retrospektivfragen verwendet, wobei der Zeitpunkt der Messung nicht dem Erhebungszeitpunkt entspricht und daher Ereignisse in der Vergangenheit abgefragt werden. Die damit verbundenen Probleme der Erinne-

rung werden mit einem prospektiven Paneldesign umgangen, da sich das erfragte Merkmal unter Berücksichtigung aller Erhebungen ergibt und dementsprechend eine hohe Datenqualität aufweist (vgl. Diekmann, 2007, S. 312).

Es wird zwischen verschiedenen Datentypen unterschieden, die sich hinsichtlich des Zeitbezuges voneinander abgrenzen. Zunächst wird zwischen Querschnittsdaten unterschieden, die Merkmale der Untersuchungseinheiten zu exakt einem Zeitpunkt liefern. Werden darüber hinaus Informationen zu verschiedenen Zeitpunkten benötigt, spricht man von Zeitreihendaten. Wesentlich umfangreicher sind dagegen Paneldaten, die eine Vielzahl von Untersuchungseinheiten zur Messung eines Merkmals zu mindestens zwei Zeitpunkten beinhalten. Kommt es dagegen auf einen besonders hohen Informationsgehalt an, bieten sich Verlaufs- oder Ereignisdaten an, da sie ein festes Zeitintervall zwischen zwei Ereignissen abbilden (vgl. ebd., S. 315).

1.2.2 KLASSIFIKATION VON VERGLEICHSSTUDIEN

Mithilfe der nachfolgenden Klassifikation nach Seidel & Prenzel (2008) lassen sich Vergleichsstudien charakterisieren.

Abbildung 1.8: Kriterien zur Klassifizierung von Vergleichsstudien

Criteria	Examples
Sample	Complete population Random sample
Scope	International National Regional
Focus assessment	Reading, mathematics, science (literacy or curriculum oriented) Basic reading and numeracy skills Specific content areas (history, literature, psychology, etc.)
Background and context	Range of included characteristics Operationalization and instruments
Target group	Age group (15-year-olds, grade 8 students, college, adults) Institutional membership (employees, unemployed, senior citizens)
Design	Cross-sectional: one-shot; trend Longitudinal: simple panel, complex panel
Initiator	Organizations (OECD, IEA, NCES) Countries States

(Seidel & Prenzel, 2008, S. 284)

Zu den charakteristischen Merkmalen einer Vergleichsstudie gehören die zu untersuchende Stichprobe (Vollerhebung, Zufallsstichprobe), die Reichweite der Studie (regional, national, international), die getestete Domäne und Ausrichtung (z. B. Lesen, Mathematik; Curriculum/Bildungsstandards, Kompetenzen), die getestete Zielgruppe (Alters-/Personengruppe), das Design der Studie (Querschnitt- und Längsschnittuntersuchung) sowie der Initiator der Studie (Organisation, Land, Staat). Die Klassifikation der renommiertesten Vergleichsstudien anhand der vorgestellten Kriterien ist in der Abbildung 1.9 exemplarisch dargestellt.

Abbildung 1.9: Klassifikation von Vergleichsstudien

	PISA	TIMSS	PIRLS	VERA	Bildungs-standards
Stichprobe	Zufallsstichprobe	Zufallsstichprobe	Zufallsstichprobe	Vollerhebung	Zufallsstichprobe
Reichweite	International	International	International	Einige Länder Deutschlands	National
Domänen	Lesen Mathematik Naturwissenschaf-ten	Mathematik Naturwissenschaf-ten	Lesen	Mathematik	Mathematik
Ausrichtung	Literacy	Curriculum	Literacy	Bildungsstandards Primarbereich	Bildungsstan-dards mittlerer Schulabschluss
Zielgruppe	15-Jährige	Klassen 4/8/12	Klassen 3/4	Klassen 3/4	Klasse 9
Design	Trend (3-jähriger Zyklus)	Trend (4-jähriger Zyklus)	Trend (5-jähriger Zyklus)		
Initiator	OECD	IEA	IEA	Land Rheinland-Pfalz	KMK
Vergleichsperspektive					
Normorien-tierter Vergleich	Internationaler Vergleich	Internationaler Vergleich	Internationaler Vergleich	Vergleich auf Landesebene	Verkoppelung mit internatio-nalem Vergleich
Kriterialer Vergleich	Kompetenzstufen	Kompetenzstufen	Kompetenzstufen	Fähigkeitsniveaus	Standards
Ipsativer Vergleich	Vergleiche über die Zeit	Vergleiche über die Zeit	Vergleiche über die Zeit	–	–
Wissenserwerb	Deskriptiv: Monitoring Benchmark Korrelationsstudie	Deskriptiv: Monitoring Benchmark Korrelationsstudie	Deskriptiv: Monitoring Benchmark Korrelationsstudie	Deskriptiv: Monitoring	Deskriptiv: Monitoring Benchmark

(Drechsel et al., 2009, S. 359)

Neben den Klassifikationskriterien werden unterschiedliche Vergleichsperspektiven ausgewiesen. Bei einer normorientierten Perspektive wird die Kompetenz zwischen Gruppen auf einer gemeinsamen Skala verglichen, um mögliche Schwächen und Stärken eines Bildungssystems aufzuzeigen. Dagegen wird innerhalb des kriterialen Vergleichs die Kompetenz von inhaltlich definierten Bildungsstandards oder -zielen gemessen. Die ipsative Perspektive dient allgemein zur Ermittlung der Kompetenzentwicklung über mehrere Messzeitpunkte (vgl. Drechsel et al., 2009, S. 357 ff.).

1.2.3 STANDARDS FÜR LÄNGSSCHNITTLICHE VERGLEICHSSTUDIEN

Im Folgenden wird ein Raster zur inhaltlichen Charakterisierung von Vergleichsstudien im Längsschnitt in Anlehnung an die Standards für empirische Untersuchungen (vgl. Seastrom, 2003; Thurgood et al., 2003; Kristen, Römmer, Müller & Kalter, 2005) entwickelt, um die Vielzahl von Studienkonzepten anhand von vier Kategorien beschreiben zu können.

1. Die Ziehung der **Stichprobe** bei Vergleichsstudien entscheidet über die Aussagekraft der Studie. Daher werden Stichproben benötigt, die sich im Sinne der Repräsentativität möglichst an der Gesamtpopulation orientieren. Zur Überprüfung von unterschiedlichen Unterrichtskonzepten bieten sich Interventions- und Kontrollgruppen zum Vergleich an.
2. Das eingesetzte **Testinstrument** basiert auf der Zielsetzung der jeweiligen Studie und kennzeichnet die methodische Erforschung von Kompetenzentwicklungen oder die Testung eines Kompetenzmodells. Hierfür werden Tests eingesetzt, die unterschiedliche Schwierigkeitsgrade abdecken und eine feste Verankerung (Linking) anhand von Items (Ankeritems) zu jedem Messzeitpunkt aufweisen können.
3. Die gewonnen Daten werden in Bezug auf die Zielsetzung der Studie mit einem umfangreichen **Auswertungsverfahren** untersucht. Dabei werden häufig die deskriptive Statistik zur Beschreibung und die induktive Statistik zur Prüfung der Daten verwendet. Des Weiteren eignet sich beispielsweise die probabilistische Testtheorie zur Bestimmung von Item-, Personen- und Modellparametern. Die genannten Verfahren werden in Kapitel 3 nähergehend dargestellt.
4. Die **Rückmeldung** der Ergebnisse erfolgt häufig durch das Berichten von Mittelwerten und Standardabweichungen bzw. -fehlern.

Querschnittstudien weisen im Gegensatz zu Längsschnittstudien im Hinblick auf diese vier Kategorien eine hohe Standardisierung zur Erfassung der Leistung in unterschiedlichen Domänen auf. Dies steht im Zusammenhang damit, dass es in der empirischen Bildungsforschung zwar groß angelegte Querschnittstudien gibt, aber keine entsprechenden Längsschnittstudien existieren, was vor allem durch den forschungsorganisatorischen Aufwand bedingt ist (vgl. Drechsel et al., 2009, S. 359). Die erste groß angelegte Längsschnittstudie in Deutschland stellt das Nationale Bildungspanel dar (vgl. Blossfeld, et al., 2009; Kapitel 6.1).

2. LEISTUNGSMESSUNG UND KOMPETENZMODELLIERUNG

Die vorliegende Arbeit befasst sich inhaltlich mit der Leistungsmessung und Kompetenzmodellierung im Längsschnitt. Die dafür notwendige Auseinandersetzung mit den Begriffen Leistung und Kompetenz erfolgt einleitend in diesem Kapitel, um nachfolgend die Kompetenzorientierung in den Bildungsstandards und die kompetenzorientierte Leistungsmessung vorzustellen. Die dazugehörigen statistischen Analyseverfahren werden detailliert in Kapitel 3 vorgestellt.

2.1 LEISTUNG UND KOMPETENZ

Zum Einstieg in die Thematik der Leistungsmessung und Kompetenzmodellierung werden die Begriffe Leistung und Kompetenz definiert und voneinander abgegrenzt.

2.1.1 LEISTUNGSBEGRIFF

Die Leistung einer Schülerin bzw. eines Schülers wird mithilfe eines Testinstruments ermittelt, das in der empirischen Bildungsforschung standardmäßig „zur Erfassung eines oder mehrerer empirisch abgrenzbarer psychologischer Merkmale mit dem Ziel einer möglichst genauen quantitativen Aussage über den Grad der individuellen Merkmalsausprägung“ (Moosbrugger & Kelava, 2012) dient. Dies bedeutet, dass die Leistung der Schülerinnen und Schüler durch die Bearbeitung der Testaufgaben in Form von Antworten sichtbar bzw. manifest[1] wird. Auf der Grundlage der manifesten Testleistung in einer Domäne wird im Sinne der Performanz (Chomsky, 1969) auf das nicht direkt beobachtbare bzw. latente[2] Kompetenzkonstrukt geschlossen (vgl. Blatt, Voss, Kowalski & Jarsinski, 2011).

[1] Ein Merkmal bzw. eine Eigenschaft, die direkt in der Beobachtung messbar ist (vgl. Gniewosz, 2011b, S. 68).
[2] Ein nicht direkt beobachtbares Merkmal einer Person, das eine hohe Komplexität aufweist und nicht direkt am Verhalten der Person erkennbar ist (vgl. Gniewosz, 2011b, S. 68).

2.1.2 KOMPETENZBEGRIFF

An dieser Stelle soll das vorherrschende Verständnis zum Kompetenzbegriff in der empirischen Bildungsforschung herausgearbeitet werden, ohne dass allerdings eine vertiefende Auseinandersetzung mit der Vielfältigkeit des Begriffs erfolgt. Eine ausführliche Darstellung und Diskussion des Kompetenzbegriffs liefern Klieme & Hartig (2007).

Zur Bestimmung des Kompetenzbegriffs im Fokus der empirischen Bildungsforschung wird das folgende Zitat von Hartig & Klieme (2006) herangezogen, da es das vorherrschende Verständnis zur Kompetenz in diesem Forschungsgebiet auf Grundlage eines von Weinert (1999) erstellten Gutachtens für die OECD zu den unterschiedlichen Kompetenzdefinitionen veranschaulicht:

> „Kompetenzen 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“ (Hartig & Klieme, 2006, S. 128).

Demnach werden Kompetenzen im Rahmen der empirischen Bildungsforschung als kognitive Leistungsdisposition definiert, womit gleichermaßen Kenntnisse, Fertigkeiten oder Routinen gemeint sind, die sich funktional auf Situationen und Anforderungen in bestimmten Kontexten beziehen. Inwiefern die definierten kontextuellen Anforderungen erfüllt werden können, lässt sich durch Kompetenz als Folge erfolgreicher Bewältigung beschreiben (vgl. Klieme et al., 2007, S. 5). Aufgrund der unterschiedlichen Kontexte entwickeln sich spezifische Kompetenzen. So werden Kompetenzen in Abhängigkeit von äußeren Faktoren beeinflusst (vgl. ebd., S. 7 f.). Die alleinige Betrachtung der kognitiven Leistungsdisposition ermöglicht die Erfassung und Untersuchung von Bildungsprozessen unabhängig von motivationalen und emotionalen Einflüssen (vgl. ebd., S. 7).

Es ist anzumerken, dass sich Kompetenzen durch ihren spezifischen Bezug von kognitiven Grundfunktionen abgrenzen (vgl. Hartig & Klieme, 2006). Somit sind Kompetenz und Intelligenz nicht gleichzusetzen, auch wenn ein hoher Zusammenhang besteht (vgl. Schweizer 2006, S. 131). Im Unterschied zur Intelligenz sind Kompetenzen veränderbar und anforderungsspezifisch. Sie knüpfen an einen bestehenden Kompetenzstand an und lösen eine Kom-

petenzentwicklung aus (vgl. Gniewosz, 2011a, S. 65). Demgegenüber wird die Intelligenz in Abgrenzung zur Kompetenz als ein generelles stabiles Konstrukt der grundlegenden kognitiven Fähigkeiten angesehen (vgl. Schweizer, 2006, S. 131).

Nachdem das Verständnis des Kompetenzbegriffs in der empirischen Bildungsforschung aufgezeigt wurde, sollen überleitend die mit den Kompetenzen verbundenen Funktionen in diesem Forschungsgebiet durch das Zitat von Hartig (2008) beschrieben werden:

> „'Kompetenzen' stellen in der empirischen Bildungsforschung zunächst 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).

Demnach werden Kompetenzen in der empirischen Bildungsforschung als theoretische Konstrukte verstanden, die sich beispielsweise durch inhaltlich abgeleitete Aufgaben operationalisieren und messen lassen. Auf dieser Grundlage können im Sinne einer evidenzbasierten Forschung empirisch geleitete Überprüfungen von Theorien bzw. Hypothesen erfolgen.

Inwiefern die inhaltliche und strukturelle Beschreibung von Kompetenzen erfolgt, wird im nächsten Schritt anhand von Kompetenzmodellen aufgezeigt.

2.1.2.1 KOMPETENZMODELLE

Die auf fachwissenschaftlichem, fachdidaktischem und pädagogisch-psychologischem Wissen basierenden Kompetenzmodelle spezifizieren Inhalte und Struktur allgemeiner Bildung. Die theoretische Anbindung an die jeweilige Fachdisziplin ermöglicht eine Transparenz im Hinblick auf Struktur und Entwicklung von Kompetenzen. Auf dieser Grundlage kann eine systematische Entwicklung von Aufgaben und Testverfahren erfolgen (vgl. Schott & Azizi Ghanbari, 2008, S. 18 ff.). Kompetenzmodelle operationalisieren demnach abstrakte Bildungsziele, indem sie konkrete messbare Aufgaben auf der Grundlage von im Idealfall theoriebasierten und empirisch überprüften Kompetenzmodellen konstruieren. Zudem legen sie wissenschaft-

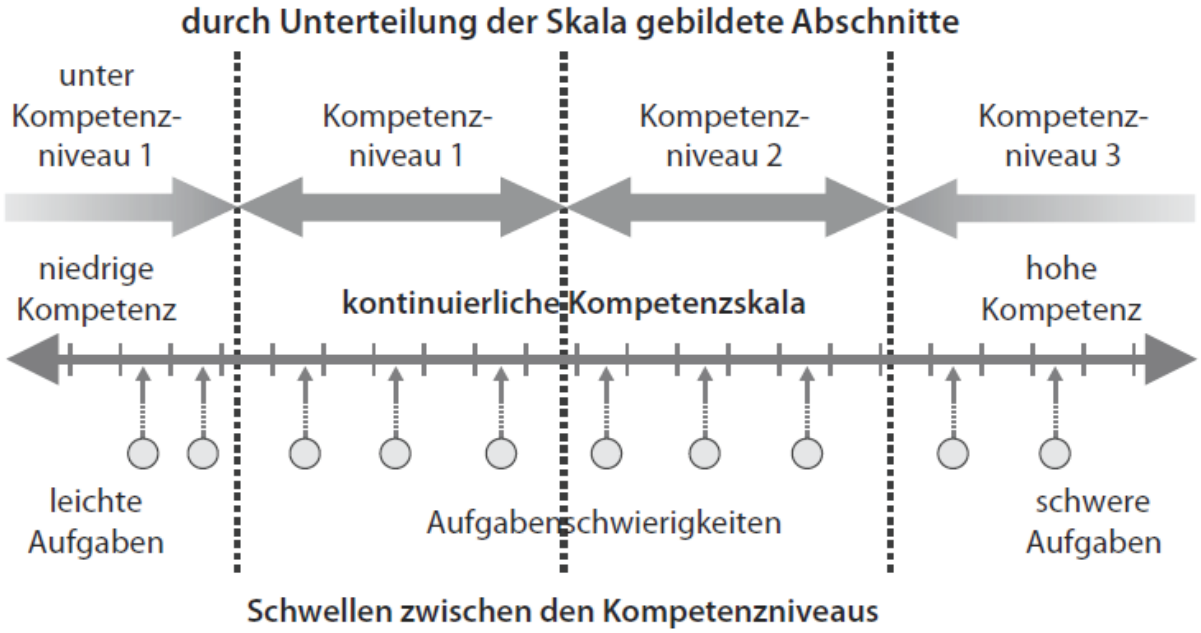
lich fundiert fest, welche Kompetenzentwicklungen von den Schülerinnen und Schülern innerhalb ihrer Schullaufbahn erwartet werden (vgl. Klieme, 2009, S. 71 ff.).

Insgesamt können zwei Arten von Kompetenzmodellen unterschieden werden: die Kompetenzniveaumodelle und die Kompetenzstrukturmodelle. Sie unterscheiden sich darin, dass bei den Niveaumodellen die konkrete inhaltliche Beschreibung von unterschiedlichen Kompetenzausprägungen vorgenommen wird, wohingegen sich Strukturmodelle mit den zu erfassenden Kompetenzdimensionen befassen (vgl. Schweizer, 2006, S. 128 ff.).

KOMPETENZNIVEAUMODELL

Innerhalb eines Kompetenzniveaumodells wird definiert, welche Anforderungen in einer Domäne bzw. in einem Schulfach gestellt werden. Sie weisen aus, welche spezifischen Anforderungen Personen mit unterschiedlich hohen Kompetenzen bewältigen können (vgl. Schweizer, 2006, S. 133 f.). Inwieweit die fachspezifischen Leistungsanforderungen durch die jeweiligen Schülerinnen und Schüler erfüllt werden können, lässt sich durch ein Kompetenzniveaumodell darstellen, indem die ermittelte individuelle Kompetenzausprägung einem Kompetenzniveau zugeordnet wird (vgl. Gniewosz, 2011a, S. 58). Dafür wird die kontinuierliche Skala der quantitativen Testwerte in Kompetenzniveaus unterteilt (vgl. Schweizer, 2006, S. 133 f.). Nach welchen Kriterien die Grenzen der Kompetenzniveaus gesetzt werden, ist wissenschaftlich nicht geklärt, da sich dieser Bereich der Kompetenzdiagnostik noch in einem recht frühen Stadium befindet und es an einschlägigen Methoden zur Definition und Prüfung von Kompetenzniveaumodellen fehlt. Die allgemeine Definition zur Setzung der Grenzen zwischen den Kompetenzniveaus werden derzeit durch die Lösungswahrscheinlichkeit der Aufgaben bestimmt, wobei ein Niveauabschnitt Aufgaben mit einer ähnlichen Schwierigkeit umfasst und daran anschließend die inhaltliche Beschreibung des Niveaus erfolgt (vgl. ebd., S. 135 f.). Das Schema eines Kompetenzniveaumodells ist in Abbildung 2.1 veranschaulicht.

Abbildung 2.1: Unterteilung einer kontinuierlichen Kompetenzskala in Kompetenzniveaus



(Schweizer, 2006, S. 135)

Bei einem Kompetenzniveaumodell sind wie oben abgebildet sowohl die Aufgaben- bzw. Itemschwierigkeiten als auch die individuellen Kompetenzausprägungen zur Bildung der Kompetenzniveaus einbezogen, indem die Aufgaben und Kompetenzen aufsteigend nach der jeweiligen Itemschwierigkeit (leicht bis schwer) sowie Kompetenzausprägung (niedrig bis hoch) auf einer Skala angeordnet werden. Demnach wird angenommen, dass das Kompetenzniveau darüber Auskunft gibt, inwieweit der Getestete Aufgaben bis zur Niveaugrenze und alle im Schwierigkeitsgrad darunterliegenden Aufgaben lösen kann.

Neben der konkreten inhaltlichen Beschreibung von unterschiedlichen Kompetenzausprägungen steht bei den Kompetenzniveaumodellen die Interpretation der durch die Kompetenzmessung gewonnenen quantitativen Testwerte für die jeweiligen Schülerinnen und Schüler im Fokus. Die zwei Möglichkeiten der Testwertinterpretation werden im folgenden Zitat beschrieben:

„Wendet man einen psychologischen Test an, so erhält man in der Regel ein numerisches Testresultat, das Auskunft über die Merkmalsausprägung der Testperson geben soll. Fragt man sich, was dieser Testwert hinsichtlich der Merkmalsausprägung aussagt, dann lässt sich diese Frage in zweierlei Weise sinnvoll beantworten: einerseits dadurch, dass der Testwert durch den Vergleich mit den Testwerten einer Bezugsgruppe interpretiert wird (normorientierte Interpretation), oder andererseits, dass eine genaue theoretische Vorstellung darüber besteht, wie der erzielte Testwert mit einem inhaltlich-psychologisch definierten Kriterium in Beziehung steht (kriteriumsorientierte Interpretation)“ (Goldhammer & Hartig, 2012, S. 174).

Die quantitativen Testwerte können demnach entweder normorientiert oder kriteriumsorientiert interpretiert werden. Bei der normorientierten Interpretation wird ein Bezug zwischen den empirisch gewonnenen Testwerten und einem Referenzwert hergestellt (vgl. Goldhammer & Hartig, 2012, S. 175 f.). So kann beispielsweise die durchschnittliche Rechtschreibleistung einer Klassenstufe oder Schule mit den Ergebnissen der internationalen Large-Scale-Assessments verglichen werden. Dagegen erfolgt die kriteriumsorientierte Testwertinterpretation unter Einbezug der konkreten Aufgaben und Anforderungen bezüglich eines spezifischen inhaltlichen Kriteriums. Dies bedeutet, dass in Abhängigkeit von den quantitativen Testwerten festgelegt wird, ab welchem Wert ein definiertes Kriterium, z. B. die erfolgreiche Zuordnung von Graphem und Phonem, angenommen werden kann (vgl. ebd., S. 182 f). Unter Zuhilfenahme der kriteriumsorientierten Testwertinterpretation ist es möglich, eine differenzierte Beschreibung der erfassten Leistung anzufertigen (vgl. Schweizer, 2006, S. 136).

Neben den Kompetenzniveaumodellen gibt es die Kompetenzstrukturmodelle, die sich mit der Dimensionalität von Kompetenzen in Form von Teilkompetenzen auseinandersetzen und nun skizziert werden (vgl. ebd., S. 132).

KOMPETENZSTRUKTURMODELL

Die Strukturmodelle weisen den korrelativen Zusammenhang zwischen Teilkompetenzen und darin enthaltenen Aufgaben aus (vgl. Gniewosz, 2011a, S. 58). Es kann nur von einer eigenständigen Dimension ausgegangen werden, wenn der korrelative Zusammenhang zwischen Teilkompetenzen ausreichend gering (Korrelation bis 0.5) ist. Dies stellt sich gerade im schulischen Bereich als Schwierigkeit dar, weil ein nachgewiesener hoher Zusammenhang (Korrelation über 0.9) bei schulischen Kompetenzen vorliegt (vgl. Schweizer, 2006, S. 133). Die Konstruktion einer Dimension bzw. einer latenten Variable kann auf unterschiedlichen Grundlagen erfolgen, beispielsweise durch die theoriegeleitete Konstruktion eines Kompetenzstrukturmodells oder die faktorenanalytische Zuordnung von hoch miteinander korrelierenden Aufgaben als Indikator für die Erfassung desselben Kompetenzmerkmals. Wie differenziert Kompetenzen erfasst werden können oder sollen, ist jedoch abhängig von zeitlichen und finanziellen Ressourcen sowie von theoretischen Faktoren (vgl. ebd., S. 132 f.).

Die differenzierte Analyse solcher Kompetenzstrukturmodelle erfolgt auf der Grundlage von mehrdimensionalen psychometrischen Modellen, wodurch die Untersuchung der erfassten Kompetenzstruktur auf Ebene der Teilkompetenzen möglich wird. Somit lassen sich differenzierte Aussagen über die Entwicklung und Förderung von Kompetenzen treffen (vgl. Klieme, et al., 2007, S. 13). Nachfolgend wird ein Blick auf die aktuelle Kompetenzforschung geworfen.

2.1.2.2 STAND DER KOMPETENZFORSCHUNG

In Deutschland dominiert die Leistungsmessung, wobei es um die Erfassung der Leistung von Schülerinnen und Schülern in unterschiedlichen Fächern geht. Mit dem Übergang ins 21. Jahrhundert hat sich Leistung im intellektuellen bzw. akademischen Sinn als das Kernthema herauskristallisiert, und zwar als Folge des großen öffentlichen Interesses an den renommierten internationalen Schulleistungsstudien (vgl. Schweizer, 2006, S. 128). In diesem Zusammenhang sind die Studien

- Trends in International Mathematics and Science Study (TIMSS)
- Programme for International Student Assessment (PISA)
- Progress in International Reading Literacy bzw. Internationale Grundschul-Lese-Untersuchung (PIRLS/IGLU)

als Large-Scale-Assessments bzw. groß angelegte Erhebungen von Leistungen der Schülerinnen und Schüler zu nennen, die Ergebnisse systematischer Leistungsmessungen für die Öffentlichkeit bereitstellen (vgl. Hartig & Klieme, 2006, S. 128). Die dabei erfasste Schülerleistung in den Domänen Lesen, Mathematik und Naturwissenschaft wird zur Modellierung spezifischer Kompetenzen genutzt (vgl. Schweizer, 2006, S. 128). Die Schulleistungsforschung setzt Standards für eine kompetenzorientierte Leistungsmessung, da sie in der Regel standardisierte Tests einsetzt, deren Entwicklung auf theoretischen Rahmenkonzepten basiert (vgl. ebd., S. 132).

Dass auf diesem Gebiet noch viel Entwicklungsarbeit geleistet werden muss, wird im Folgenden anhand des von der Deutschen Forschungsgemeinschaft (DFG) seit 2007 geförderten Schwerpunktprogramms „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen" (Laufzeit 2007-2013) aufgezeigt. Es wurde 2005 von Eckhard Klieme (Deutsches Institut für Internationale Pädagogische Forschung, DIPF) und Detlev Leutner (Universität Duisburg-Essen) beantragt (vgl. Jude & Klieme, 2008, S. 9). Der Antrag (Klieme & Leutner, 2006a) bietet die bis dahin umfassendste und strukturierteste Darstellung des aktuellen Forschungstandes zur Kompetenzdiagnostik im Rahmen der empirischen Bildungsforschung (vgl. Schott & Azizi Ghanbari, 2008, S. 16). Das dort zugrunde gelegte Verständnis von Kompetenz als „*kontextspezifische kognitive Leistungsdispositionen*, die sich funktional auf Situationen und Anforderungen in bestimmten *Domänen* bezieht" (Klieme & Leutner, 2006a, S. 4 f.), wurde bereits in Kapitel 2.1.2 erläutert.

Im Rahmen des Schwerpunktprogrammes (SPP) sollen als wichtiger Beitrag zur Bildungsforschung die international verbreiteten Arbeiten aus Deutschland zu den Grundlagen und Methoden, die der Erfassung der Outputfaktoren des Bildungssystems dienen, gesichtet und zielgerichtet zusammengefasst werden (vgl. ebd., S. 1):

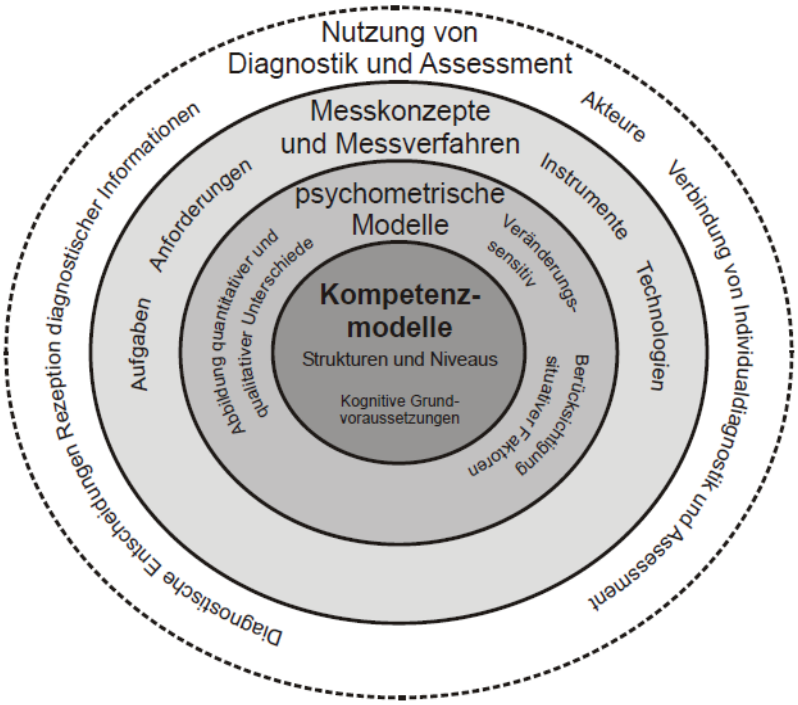
Diese Zusammenfassung zeigt, dass es bislang weder national noch international zufriedenstellend gelungen ist, theoretische Kompetenzmodelle für kognitive Kompetenzen mit psychometrischen Modellen und Messverfahren adäquat zu verknüpfen und die Kompetenzmodelle als Grundlage für pädagogische Entscheidungen zu nutzen. Mit der Durchführung des SPP sollen die wissenschaftlichen Voraussetzungen für eine empirisch geleitete Förder-, Platzierungs- und Auswahlentscheidung, Benotung und Zertifizierung von Schülerinnen und Schülern sowie für die Evaluation von Bildungsangeboten und für die Evaluation des Bildungssys-

tems geschaffen werden. Mit dem Fokus auf der Grundlagenforschung wird im SPP die Nutzung von Kompetenzmodellen in unterschiedlichen pädagogischen Entscheidungssituationen untersucht. Vor allem geht es bei diesem Programm um die interdisziplinäre Kooperation bei der Modellierung und Erfassung von Kompetenzen im Bildungsbereich bezogen auf konkrete Lern- und Handlungsbereiche. Besonders häufig erfolgt dabei die Erfassung der Lesekompetenz, der mathematischen und naturwissenschaftlichen Kompetenz sowie der Problemlösekompetenz. An diesen Vorarbeiten möchte das SPP anknüpfen und mit deren Erkenntnissen neue Wege zur Messung von Kompetenzen auf Basis von kognitiven Prozessmodellen entwickeln (vgl. Klieme & Leutner, 2006b, S. 876 ff.).

„Das angestrebte Schwerpunktprogramm soll dementsprechend als integratives Forschungsprogramm kognitiv orientierte Experten auf den Gebieten einzelner bereichsspezifischer Kompetenzen mit Experten auf dem Gebiet des Messens und Testens zusammenbringen. Ziel ist es, Kompetenzstruktur- und Kompetenzentwicklungsmodelle zu erarbeiten und empirisch zu prüfen, anhand derer sich valide und faire Messinstrumente auf zwei Ebenen konstruieren lassen: zum einen auf der Ebene der Förderung individueller Lernprozesse (im Sinne einer individuellen „Kompetenzdiagnostik"), zum anderen auf der Ebene von Untersuchungen zum Monitoring von Bildungsinstitutionen und Bildungssystemen („Assessment"). Abgerundet wird das Programm durch Forschung zur Nutzung von kompetenzbezogenen Messinstrumenten in unterschiedlichen pädagogischen Entscheidungskontexten" (Klieme & Leutner, 2006a, S. 3 f.).

Durch die Zusammenführung von Experten aus einzelnen Domänen werden im Rahmen des Schwerpunktprogramms die Erarbeitung und empirische Überprüfung von Kompetenzstruktur- und Kompetenzentwicklungsmodellen sowie die darauf basierende Konstruktion valider und fairer Messinstrumente vorangebracht. Ergänzt wird das Vorhaben durch einen Praxisbezug, indem das kompetenzbezogene Wissen für pädagogische Entscheidungssituationen genutzt wird. Dadurch entsteht eine neue Qualität für die Messung von Lernvoraussetzungen und -ergebnissen (vgl. Klieme & Leutner, 2006a, S. 1). Die zentrale Fragestellung des Schwerpunktprogramms umfasst vier aufeinander aufbauende Punkte, die in Abbildung 2.2 dargestellt sind.

Abbildung 2.2: Arbeitsbereiche des Schwerpunktprogramms



(Klieme & Leutner, 2006a, S. 6)

Im Kern zielt das Forschungsvorhaben auf die Entwicklung theoretischer Kompetenzmodelle (1) ab, die die kognitive Grundvoraussetzung der Testperson berücksichtigt und die Strukturen der Kompetenz sowie deren Niveaus definiert. Für die Kompetenzmodellierung werden psychometrische Messmodelle (2) eingesetzt, um quantitative sowie qualitative Kompetenzunterschiede und -entwicklungen in Abhängigkeit von situativen Faktoren zu erfassen. Zur domänenspezifischen Erfassung der Kompetenzen werden Messkonzepte und -verfahren (3) benötigt. Der letzte Bereich „Nutzung von Diagnostik und Assessment" (4) ermöglicht die Nutzung der gewonnen Informationen über die jeweiligen Akteure, um z. B. die Notenvergabe der Lehrkräfte zu validieren (vgl. Klieme & Leutner, 2006b, S. 882).

Hieran knüpfen vier Leitfragen an, die entsprechend das übergeordnete Forschungsziel festlegen, und zwar neue wissenschaftliche Wege bei der theoretischen Konstruktion und der empirischen Validierung von Kompetenzmodellen zu gehen, die sich gleichermaßen in psychometrische Modelle und in empirische Messverfahren überführen lassen. Die vier zugrundeliegenden Leitfragen werden im Folgenden erläutert (Klieme & Leutner, 2006a, S. 6 ff.):

1. *„Wie lassen sich Kompetenzen, unter Berücksichtigung ihres Bezugs auf Anforderungen in spezifischen Situationen, angemessen kognitiv modellieren?"*

Hiermit ist im Sinne des SPP die Auseinandersetzung mit der Kompetenzstruktur und den Kompetenzniveaus gemeint, um umfassende Kompetenzbeschreibungen zu erarbeiten. Diese müssen den kontextspezifischen Anforderungen von Kompetenzen und den jeweiligen domänenspezifischen Eigenschaften gerecht werden, um theoriegeleitete Aufgaben entwickeln zu können, die kognitive Entwicklungsmerkmale, Kompetenzniveaus und die längsschnittliche Kompetenzentwicklung berücksichtigen.

2. *„Wie lassen sich theoretische Kompetenzmodelle in psychometrischen Modellen abbilden, um die Kompetenzkonstrukte einer differenziellen Erfassung zugänglich zu machen?"*

Die Psychometrie liefert komplexe Messmodelle zur Abbildung von kontextualisierten Item- und Personenmerkmalen, kategorialen Unterscheidungen und Veränderungen. Auf dieser Grundlage soll die Forschung beim SPP fortgeführt werden und für eine Verzahnung der weiteren drei Arbeitsbereiche sorgen. So sollen die kognitiven Kompetenzmodelle (Bereich 1), die konkreten Messkonzepte und -verfahren (Bereich 2) und deren Nutzung für Diagnostik und Assessment (Bereich 4) interagieren.

3. *„Wie lassen sich Kompetenzmodelle und darauf basierende psychometrische Modelle in konkrete empirische Messverfahren übertragen?"*

Zur Erfassung der Kompetenz in den unterschiedlichen Domänen strebt das SPP die Konstruktion von neuen Messkonzepten an. So lassen sich Kompetenzmodelle in psychometrische Modelle und konkrete Messverfahren überführen, die sowohl der Grundlagenforschung als auch der kontextuellen Anwendung dienen.

4. *„Welche Arten von Informationen aus Kompetenzmessungen können von Akteuren im Bildungswesen auf welche Weise genutzt werden?"*

Inwiefern die Bildungspraxis und -politik Nutzen aus den Studienergebnissen zu Diagnostik und Assessment ziehen kann, stellt ein weiteres Anliegen des SPP dar. Auch wenn dieses Teilziel zurzeit noch eine untergeordnete Rolle spielt, erwarten die Forscher, dass die Forschungsergebnisse dazu beitragen, die Validität pädagogischer Entscheidungen zu steigern (vgl. Klieme & Leutner, 2006b, S. 883 ff.).

2.1.3 ABGRENZUNG DER BEGRIFFE LEISTUNG UND KOMPETENZ

Im Zusammenhang mit der Kompetenzorientierung in der empirischen Bildungsforschung und insbesondere durch die Etablierung von Bildungsstandards nimmt der Gebrauch des Kompetenzbegriffs exponentiell zu. Dabei ist festzustellen, dass der Begriff in einem weiteren Sinn gleichbedeutend mit Leistung verwendet wird. Die LOGIK- und SCHOLASTIK-Studie (Kapitel 4.1.1) zielt z. B. auf die „Untersuchung der Entwicklung von Lese- und Rechtschreibkompetenzen im Verlauf der Grundschulzeit" (Schneider, Stefanek & Dotzler, 1997, S. 127), ohne dass den Studien ein Kompetenzmodell zugrunde gelegt ist.

Aus wissenschaftlicher Sicht müssen diese Begriffe jedoch konkretisiert und voneinander abgegrenzt werden. Unter Leistung ist das durch einen Test gemessene Ergebnis zu verstehen, unter Kompetenz die dieser Leistung zugrundeliegenden Kenntnisse, Fähigkeiten und Fertigkeiten. Voss (2009) nennt als Voraussetzungen für eine kompetenzorientierte Leistungsmessung, dass die zur Leistungsmessung eingesetzten Tests auf einem theoretischen Rahmenkonzept basieren und dass dieses Konzept mit den Testdaten durch eine Kompetenzmodellierung empirisch validiert wird. Klieme (2009) hebt die Bedeutung der Fächer und Fachdidaktiken bei der Entwicklung von Kompetenzmodellen hervor, die für die Entwicklung des theoretischen Rahmenkonzepts zuständig sind.

> „Eine Konsequenz ist, dass konkrete Ausformulierungen und Operationalisierungen des Kompetenzbegriffs zunächst in den Domänen bzw. Fächern zu erfolgen haben. Daraus begründet sich weiterhin die Notwendigkeit, bei der Entwicklung von Kompetenzmodellen auf dem Theorie- und Erkenntnisstand der Fachdidaktiken aufzubauen" (Klieme, 2009, S. 75).

In Anlehnung an Blatt et al. (2011) wird im Rahmen dieser Arbeit der Begriff Leistung in Bezug auf die „Datenerhebung" und der Begriff Kompetenz im Zusammenhang mit der „Datenanalyse und -modellierung" benutzt (S. 230).

2.1.4 EINFLUSSFAKTOREN AUF DIE KOMPETENZ

Um sich vertiefend mit der Kompetenz auseinandersetzen zu können, werden in der empirischen Bildungsforschung neben der Leistung auch Hintergrundinformationen zu den geteste-

ten Schülerinnen und Schülern erhoben. Diese Informationen werden über Fragebögen oder Interviews gewonnen und liefern Daten, um sogenannte Einflussfaktoren zu erfassen, mit denen sich Unterschiede und Effekte in Bezug auf die Entwicklung bzw. Veränderung der Kompetenz bzw. der Kompetenzstruktur erklären lassen. In den international vergleichenden Schulleistungsstudien in Deutschland (vgl. Abbildung 1.9) hat sich ein Kernbereich von Faktoren herausgebildet, die obligatorisch erhoben werden (vgl. Bos, Tarelli, Bremerich-Vos, Schwippert 2012; Bos, Wendt, Köller & Selter, 2012; Prenzel, Sälzer, Klieme & Köller, 2013). Dazu zählen:

- Geschlecht
- Alter
- Sozioökonomischer Status
- Migrationshintergrund
- Schulform

In Bezug auf das Geschlecht zeigen sich in der nationalen und internationalen Forschung bereits in der frühen Schullaufbahn Kompetenzunterschiede zwischen Jungen und Mädchen, wie es beispielsweise die jüngsten IGLU- und TIMS-Studien aus dem Jahr 2011 erneut nachgewiesen haben (vgl. Bos, Bremerich-Vos, Tarelli & Valtin, 2012, S. 91 ff.; Brehl, Wendt & Bos, 2012, S. 203 ff.). Das Alter der Schülerinnen und Schüler dient als Anhaltspunkt für den Zeitpunkt der Einschulung und kann in Form des durchschnittlichen Alters als Einflussfaktor untersucht werden um herauszufinden, ob sich altersspezifische Unterschiede bei der Kompetenz erkennen lassen (vgl. Tarelli, Lankes, Drossel & Gegenfurtner, 2012, S. 145 f.). In einem engen Zusammenhang zur Leistung der Schülerinnen und Schüler steht der sozioökonomische Status der Familie. Um ihn zu erheben, wird in Anlehnung an Bourdieu (1983) und Coleman (1988) das ökonomische, kulturelle und soziale Kapital erfasst, indem beispielsweise für das kulturelle Kapital die Anzahl der Bücher im Haushalt erfragt werden (vgl. Wendt, Stubbe & Schwippert, 2012, S .175 ff.; Stubbe, Tarelli & Wendt, 2012, S. 231 ff.). Ob ein Migrationshintergrund bei den Schülerinnen und Schülern vorliegt, wird über die gesprochene Sprache im Elternhaus und/oder das Geburtsland der Eltern erfragt. So zeigen sich bei Schülerinnen und Schülern mit Migrationshintergrund z. B. positive Effekte auf die Leistung, wenn Deutsch in der Familie gesprochen wird, im Unterschied zu Schülerinnen und Schülern, die in ihrem Elternhaus die Herkunftssprache sprechen (vgl. Tarelli, Schwippert & Stubbe, 2012, S. 247 ff.). Mit der Kontrolle der Schulform lässt sich die Kompetenz in Abhängigkeit von der Schulart be-

trachten, um auf diese Weise durch das Schulsystem bedingte Unterschiede bei der Ausprägung der Kompetenz festzustellen. Diesbezüglich zeigt sich z. B. bei PISA 2012, dass sich die Schulformen zwar bei der festgestellten Kompetenz überlappen, aber auch, dass es deutliche Unterschiede zwischen den Extremen Hauptschulen und Gymnasien gibt und dass die Schulform mit über den Bildungserfolg und die gesellschaftliche Teilhabe der Schülerinnen und Schüler bestimmt (vgl. Sälzer, Reiss, Schiepe-Tiska, Prenzel & Heinze, 2013, S. 86 ff.)

2.2 KOMPETENZORIENTIERUNG IN BILDUNGSSTANDARDS

Bildungsstandards stellen eine Möglichkeit dar, Kompetenz zu operationalisieren, indem kompetenzorientierte Bildungs- bzw. Lernziele definiert werden und deren Erreichen empirisch untersucht werden kann. Im Gegensatz zur kompetenzorientierten Leistungsmessung im wissenschaftlichen Sinn (Kapitel 2.3) wird bei den Bildungsstandards zwar eine kompetenzorientierte Leistungsmessung anhand der an Kompetenzbereichen orientierten Bildungs- bzw. Lernziele vorgenommen, aber es wird auf eine Kompetenzmodellierung auf Basis eines zugrundeliegenden theoretischen Kompetenzmodells und der Validierung des Kompetenzmodells verzichtet. Welche Erwartungen und Funktionen mit den Bildungsstandards verknüpft sind, wird in Verbindung mit der Überprüfung der definierten Ziele durch landesbezogene Vergleichsarbeiten (VERA) aufgezeigt.

Die Einführung von nationalen Bildungsstandards in der Primarstufe (4. Klassenstufe) und in der Sekundarstufe (mittlerer Abschluss in Klassenstufe 9. bzw. 10) (vgl. Klieme et al., 2003) dient nach Auffassung der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Kultusministerkonferenz, KMK) zur Spezifizierung des fachlichen Lernens im Kontext allgemeiner Bildungsziele (vgl. Klieme, 2009, S. 11 ff.). Mit den Bildungsstandards hat der Kompetenzbegriff die Funktion erhalten, messbare Ziele des Bildungssystems zu definieren und zu charakterisieren (vgl. Hartig & Klieme, 2006, S. 128). Die Operationalisierung der Bildungsziele geschieht nach Klieme (2009) folgendermaßen:

„Nationale Bildungsstandards formulieren verbindliche Anforderungen an das Lehren und Lernen in der Schule. Sie stellen damit innerhalb der Gesamtheit der Anstrengungen zur Sicherung und Steigerung der Qualität schulischer Arbeit ein zentrales Gelenkstück dar. Bildungsstandards benennen präzise, verständlich und fokussiert die wesentlichen Ziele der pädagogischen Arbeit, ausgedrückt als erwünschte Lernergebnisse der Schülerinnen und Schüler. Damit konkretisieren sie den Bildungsauftrag, den Schulen zu erfüllen haben" (Klieme, 2009, S. 9).

Mit den Bildungsstandards wird festgelegt, über welche Kompetenzen Schülerinnen und Schüler in der jeweiligen Jahrgangsstufe verfügen sollen, indem die jeweiligen Kompetenzanforderungen im Verlauf der Schullaufbahn definiert werden und eine für alle verbindliche Konzentration auf Kernkompetenzen erfolgt. Die in den Bildungsstandards festgelegten Bildungsziele der Kernfächer werden mit domänenspezifischen Messverfahren erfasst (vgl. Klieme et al., 2007, S. 10). Auf dieser Grundlage bieten die Bildungsstandards die Möglichkeit, den Output des Bildungssystems durch ein regelmäßiges Bildungsmonitoring oder Schulevaluationen zu kontrollieren, womit die Schulentwicklung durch die dabei gewonnenen Ergebnisse profitieren kann (vgl. Klieme, 2009, S. 9 ff.).

Neben den durch Bildungsstandards definierten Kompetenzen werden auch sogenannte fachübergreifende Kompetenzen bzw. Schlüsselkompetenzen erfasst (vgl. Hartig & Jude, 2007, S. 25).

2.2.1 VERGLEICHSARBEITEN – VERA

Zur Überprüfung der Bildungsstandards werden landesweite Vergleichsarbeiten (VERA)[3] eingesetzt (vgl. KMK, 2012). Dazu werden bundesweit einheitliche schriftliche Tests in den Fächern Deutsch und Mathematik zur Untersuchung der Schülerkompetenzen in der 3. und 8. Klassenstufe (VERA-3 und VERA-8) zum Bildungsmonitoring durchgeführt, wobei die Länder Kompetenzbereiche auswählen können. Daneben gibt es landesspezifische Vergleichsarbei-

[3] In manchen Ländern sind die Vergleichsarbeiten unter anderen Bezeichnungen bekannt (vgl. www.iqb.hu-berlin.de/vera).

ten, z. B. in der Klassenstufe 6 (vgl. Blatt, Ramm & Voss, 2009). Mit dem bundesweit einheitlichen Bildungsmonitoring soll die Kompetenzorientierung im Bildungssystem etabliert werden.

So wird bei VERA-3 verpflichtend einmal im Jahr mindestens eines der Fächer Deutsch oder Mathematik pro Land getestet. Die Fächer umfassen mehrere Kompetenzbereiche. Für das Fach Deutsch wird beispielsweise differenziert zwischen „Lesen", „Zuhören", „Orthografie" sowie „Sprache und Sprachgebrauch". Zukünftig soll der Kompetenzbereich „Schreiben" mit aufgenommen werden. Bei der Mathematik werden im Vorfeld zwei von fünf möglichen Kompetenzbereichen wie „Zahlen und Operationen" oder „Daten, Häufigkeiten und Wahrscheinlichkeiten" für die Testung ausgewählt.

Bei VERA-8 werden ebenfalls in mindestens einem der Fächer Deutsch und Mathematik schriftliche Tests durchgeführt. Im Fach Deutsch sind die Kompetenzbereiche zu VERA-3 bundesweit identisch. Im Fall der Mathematik werden alle fünf Kompetenzbereiche getestet. Hinzu kommt die Erfassung der Kompetenz in der ersten Fremdsprache, wobei mindestens das „Lese- und Hörverstehen" in den Fächern Englisch oder Französisch erfasst wird.

Den Ländern steht es frei, mehr als ein Fach bzw. einen Kompetenzbereich zu testen. Für jede Erhebungswelle werden neue Aufgaben durch das Institut zur Qualitätssicherung im Bildungswesen (IQB) entwickelt. Zu den Aufgaben gibt es Auswertungsvorlagen mit formalen, sprachlichen, strukturellen und inhaltlichen Kriterien. Weiterhin gibt es zu den Aufgaben sogenannte didaktische Handreichungen und Kommentierungen. In den didaktischen Handreichungen werden die Kompetenzbereiche definiert. Die didaktische Kommentierung dient zur Reflexion der Aufgabenmerkmale und des diagnostischen Potenzials und liefert fachdidaktische Anregungen für die anschließende Weiterarbeit und Förderung.

2.3 KOMPETENZORIENTIERTE LEISTUNGSMESSUNG

Dieses Kapitel befasst sich mit der kompetenzorientierten Leistungsmessung im Rahmen der empirischen Bildungsforschung. Dazu zählt die Kompetenzmessung, bei der die Leistung der Schülerinnen und Schüler anhand eines theoretisch begründeten Kompetenzmodells erfasst wird. Des Weiteren umfasst die kompetenzorientierte Leistungsmessung die Modellierung der Kompetenz auf Grundlage der probabilistischen Testtheorie. Somit werden im Folgenden die Kompetenzmessung und Kompetenzmodellierung thematisiert.

2.3.1 KOMPETENZMESSUNG

Mit der empirischen Messung von Kompetenzen sind konkrete Ziele verbunden: Durch Kompetenzmessungen erfolgt die Evaluation und das Monitoring des Bildungssystems zur Qualitätssicherung. Ebenso wird die Kompetenzmessung zur Beurteilung von Schülerinnen und Schülern genutzt, um daran anschließend gezielte Fördermaßnahmen abzuleiten. Gleichermaßen leistet die Kompetenzmessung einen Beitrag für die Grundlagenforschung, wenn es um die Erklärung von Bedingungen und Fördermöglichkeiten der Kompetenzentwicklung geht (vgl. Hartig & Jude, 2007, S. 17).

Im Unterschied zu den Vertretern einer möglichst umfassenden Kompetenzerfassung schlägt Weinert (2001) in Bezug auf die vergleichende Kompetenzmessung eine Konzentration auf die fachbezogenen kognitiven Leistungsdimensionen vor (vgl. Jude & Klieme, 2008, S. 10).

Wie Tests in der empirischen Bildungsforschung eingesetzt werden, beschreibt das folgende Zitat.

> „In der empirischen Bildungsforschung konzentriert sich die Verwendung von Tests auf Merkmale von Personen, die im Zusammenhang zu Bildung und Lernen stehen. Zwei Varianten von Tests sind in der empirischen Bildungsforschung besonders wichtig: Leistungs- und Persönlichkeitstest“ (Gniewosz, 2011b, S. 70).

Die Messung von Merkmalen erfolgt in der empirischen Bildungsforschung mit psychometrischen Leistungs- und Persönlichkeitstests. Bei einem Leistungstest kann für jede Aufgabe unterschieden werden, ob es sich bei der Antwort um eine richtige oder falsche Angabe handelt. Die Anzahl der richtig gelösten Aufgaben einer Person bestimmt die gemessene individuelle Leistung. Mit dem individuellen Leistungswert lässt sich im Vergleich zu einer Normstichprobe der Leistungsstand der getesteten Person bestimmen, also wie stark die Leistung des getesteten Merkmals bei der Person ausgeprägt ist und sich von der Normstichprobe unterscheidet. Innerhalb der Leistungstests wird weitergehend zwischen sogenannten Power- und Speedtests unterschieden. Bei den Powertests gibt es keine zeitliche Einschränkung bei der Bearbeitung des Tests. Durch die ansteigende Itemschwierigkeit ist eine Person nur in der Lage, den Test bis zu der Aufgabe zu bearbeiten, für die die individuelle Leistungsfähigkeit ausreicht. Die Speedtests unterliegen einem Zeitlimit, weshalb es möglicherweise nicht allen

Testpersonen gelingt, alle Aufgaben zu bearbeiten. Persönlichkeitstests dienen der Messung individueller Merkmale, beispielsweise bezogen auf das Interesse und die Einstellung zu dem zu erfassenden Merkmal (vgl. Gniewosz, 2011b, S. 70).

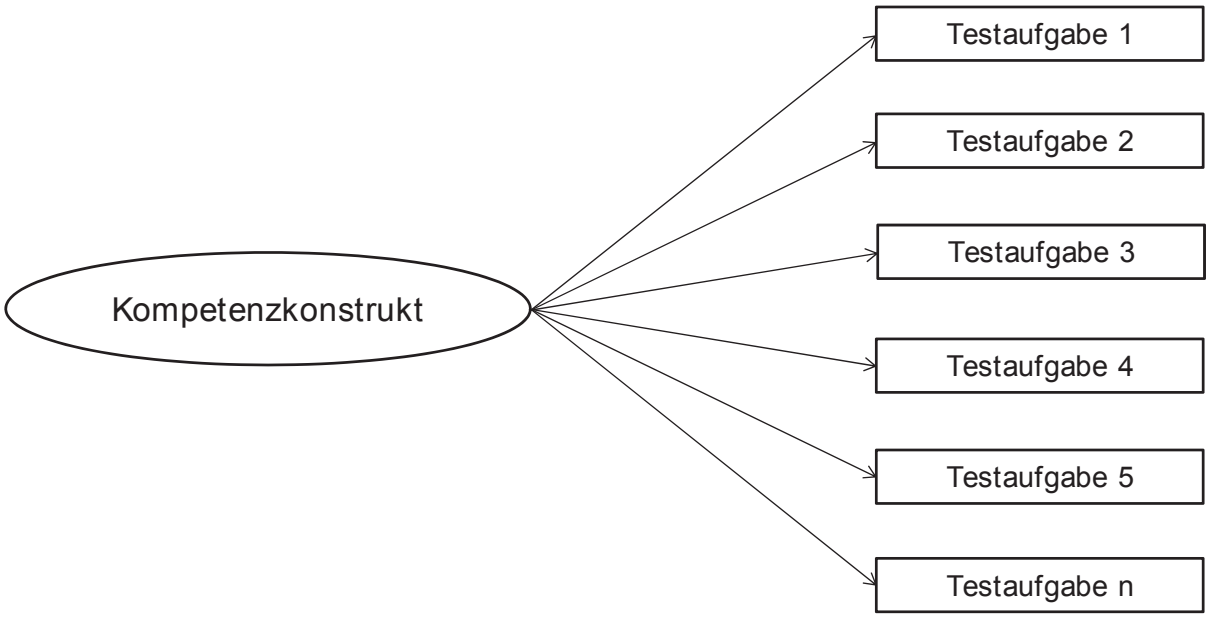
Mit psychometrischen Tests können drei Arten von Aussagen getroffen werden: Mithilfe psychometrischer Tests gewonnene Ergebnisse erlauben erstens die Aussage, wie zuverlässig ein latentes Merkmal gemessen werden kann. Sie erlauben zweitens Angaben über individuelle Merkmalsausprägungen. Drittens kann die Verteilung des Merkmals hinsichtlich des Mittelwerts und der Varianz in einer Gruppe von Personen bestimmt werden (vgl. ebd., S 68 f.). Damit diese Aussagen möglich sind, müssen folgende Voraussetzungen erfüllt sein: Die Tests sollten im Sinne einer Normierung möglichst bei einer repräsentativen Normstichprobe eingesetzt werden, sodass Merkmale der Stichprobe mit der Grundgesamtheit verglichen werden können. Zudem muss gewährleistet sein, dass das zu erfassende Merkmal durch ausreichend viele Aufgaben operationalisiert ist und so Rückschlüsse auf die individuelle Merkmalsausprägung möglich sind. Letztlich müssen die Tests Gütekriterien als Qualitätsstandards erfüllen, damit davon ausgegangen werden kann, dass die Messung sich tatsächlich auf das zu erfassende Merkmal bezieht (vgl. ebd., S 69).

Die an die empirischen Messverfahren gestellten qualitativen Ansprüche werden anhand der drei klassischen Gütekriterien *Objektivität, Reliabilität* und *Validität* überprüft. Das Zusammenspiel der Gütekriterien wird durch die Objektivität bestimmt, da sie ein notwendiges Kriterium für die Betrachtung der Reliabilität und Validität darstellt. Nur wenn die Gütekriterien eingehalten werden, können die Ergebnisse der Kompetenzmessung zur Beantwortung von Forschungsfragen oder für Aussagen über das Bildungssystem genutzt werden (vgl. Hartig & Jude, 2007, S. 19). Wenn die Messung, Auswertung und Interpretation standardisiert und dokumentarisiert abläuft, ist davon auszugehen, dass die Testung nicht zur Beeinflussung des Testergebnisses einer Person beiträgt und die Ergebnisse im Sinne der Objektivität allein auf deren individuelle Leistungsdisposition zurückgehen. Ob das Testinstrument das Kompetenzkonstrukt messen kann, wird anhand der Reliabilität eines Tests abgeschätzt. Dazu wird der korrelative Zusammenhang der Testaufgaben herangezogen, der als interne Konsistenz bezeichnet wird und abhängig von der Höhe der Korrelation auf die Messgenauigkeit des Tests schließen lässt. Als wichtigstes Gütekriterium wird die Validität zur Beurteilung von Messverfahren angesehen, da sie die Gültigkeit des zu untersuchenden Kompetenzmerkmals eines Tests sicherstellt. Da die Validität auch über die Zuverlässigkeit ergebnisorientierter Aussagen

entscheidet, existieren eine Reihe von Validitätsuntersuchungen zur Bestimmung der Aussagengültigkeit (Borsboom, Mellenbergh & van Herrden, 2004). Beispielsweise wird bei der für die weitere Ausführung dieser Arbeit zutreffenden Inhaltsvalidität geprüft, inwiefern die Testaufgaben das zu untersuchende Kompetenzkonstrukt repräsentieren. Dies kann nicht alleine aus der empirischen Messung des Konstrukts gefolgert werden, sondern bedarf in Bezug auf curriculare Lernziele einer fachdidaktischen Beurteilung. Inwieweit ein Test in der Lage ist, ein Kompetenzkonstrukt zu erfassen, wird also durch Expertenurteile festgelegt (vgl. Hartig & Jude, 2007, S. 23). Für die Bestimmung der Gütekriterien können Modelle der Item-Response-Theory (IRT) (Kapitel 3.1) verwendet werden (vgl. ebd., S. 21 f.).

Da das zu messende Kompetenzkonstrukt durch die Auswahl der Testaufgaben festgelegt wird, bedarf es einer wohlüberlegten Strategie zur Testkonstruktion. Zur Auswahl stehen hierfür die externale, deduktive oder induktive Konstruktion. Beispielsweise wird im Rahmen der deduktiven Testentwicklung idealerweise das Kompetenzkonstrukt durch ein theoretisches Modell beschrieben, woraus sich fundierte Begründungen für die Ableitung von Indikatoren bzw. Aufgaben ergeben. Dies ermöglicht es, theoriebegründete Annahmen über den Zusammenhang zwischen der zu untersuchenden Kompetenz und der Aufgabenlösung aufzustellen, also die Wahrscheinlichkeit zu bestimmen, mit der eine weniger kompetente bzw. kompetente Person eine Aufgabe löst. Aufgrund der theoretischen Fundierung stellt die deduktive Testentwicklung einen sowohl gleichermaßen anspruchsvollen als auch vielversprechenden Ansatz im Bereich der Kompetenzdiagnostik dar, um Annahmen über die Beschreibung und Unterscheidung personenbezogener Kompetenzen zu formulieren. Die aus dem theoretischen Konstrukt abgeleiteten messbaren Indikatoren bilden schließlich das Messinstrument (vgl. Hartig & Jude, 2007, S. 26 ff.). Die gemessenen individuellen Testwerte werden als latentes Merkmal abgebildet, stellen also eine nicht direkt beobachtbare Größe dar (vgl. ebd., S. 32), wie es in der nachfolgenden Abbildung 2.3 veranschaulicht ist.

Abbildung 2.3: Messmodell zur Erfassung einer Kompetenz



(Modifizierte Abbildung nach Voss, 2009, S. 15)

Die konkreten Testwörter bzw. Testaufgaben repräsentieren also das Messmodell eines nicht direkt beobachtbaren Kompetenzkonstrukts. Sofern die Testaufgaben geeignet sind, das zu erfassende Kompetenzkonstrukt reliabel und in kohärenter Weise zu messen, ist die Lösungswahrscheinlichkeit der Aufgaben allein von der jeweiligen Leistung der Schülerinnen und Schüler abhängig. Hierbei liegt der Fokus nicht auf den einzelnen Aufgaben, sondern der gesamte Test dient als Grundlage für die kompetenzorientierte Leistungsmessung, wodurch mögliche Messfehler einzelner Aufgaben relativiert werden.

Die aus der kompetenzorientierten Leistungsmessung gewonnenen Daten werden für die Modellierung der Kompetenz und der Validierung des Kompetenzmodells verwendet, was im Folgenden vorgestellt wird.

2.3.2 KOMPETENZMODELLIERUNG

Die mithilfe eines Testinstruments erfassten Leistungsdaten der Schülerinnen und Schüler werden zur Modellierung der Kompetenz genutzt. Zur Modellierung der Kompetenz werden unterschiedliche Verfahren kombiniert, die im Folgenden beispielhaft dargestellt werden.

Die bei der Bearbeitung eines Tests erfolgte Beantwortung der Testaufgaben durch die Schülerinnen und Schüler wird in ein auswertbares Format übertragen. Die Kompetenzmodellierung erfordert daher zunächst eine qualitative Auswertung der Antworten, um deren Qualität zu ermitteln. Zu diesem Zweck werden die Antworten exakt übertragen bzw. transkribiert und durch eine manuelle oder automatische Kodierung nach richtigen oder falschen Antworten bewertet. Für diesen Vorgang sind entweder besonders geschulte Kodierer oder eine speziell entwickelte Software notwendig. Die auf diesem Weg übertragene manifeste Leistung führt zu einem Datensatz als Grundlage für die Modellierung der latenten Kompetenz.

Aufgrund der Verwendung von Verfahren der Item-Response-Theory (IRT) zum Skalieren der Leistung (Embretson & Reise, 2000; Rost, 2004) spricht man in diesem Zusammenhang von einer psychometrischen Modellierung der Schülerkompetenzen (vgl. Voss, 2009, S. 14). Zur Beschreibung des Kompetenzkonstrukts werden mittels der IRT sogenannte Item- und Personenparameter geschätzt, die in diesem Fall die Schwierigkeiten der Testaufgaben und die Personenfähigkeiten repräsentieren. Je nach theoretischer Grundlage des Kompetenzkonstrukts kann sich die Darstellung der Kompetenz ein- oder mehrdimensional gestalten. Bei einem eindimensionalen Kompetenzkonstrukt handelt es sich um ein einzelnes kontinuierliches latentes Merkmal, wohingegen bei einem mehrdimensionalen Kompetenzkonstrukt einzelne Teilkompetenzen im Sinne eines differenziellen Kompetenzmodells unterschieden werden. Dies ermöglicht differenzierte Aussagen über die Teilkompetenzen einer Kompetenz, um z. B. die Zusammenhänge zwischen den zugrundeliegenden Teilkompetenzen in Form von separaten latenten Dimensionen zu untersuchen (vgl. Hartig & Jude, 2007, S. 33). Die differenzierte Betrachtung bzw. Darstellung des Kompetenzkonstrukts erfordert allerdings eine ausreichend große Stichprobe und eine entsprechende theoretische Begründung, um eine valide Umsetzung der Teilkompetenzen zu gewährleisten. So können z. B. komplexe Kompetenzmodelle mit einzelnen Teilkompetenzen berücksichtigt und mehrdimensional abgebildet werden. Inwiefern eine Passung zwischen dem theoretischen und empirischen Modell vorliegt, wird mit der Skalierungssoftware ermittelt. Dazu werden informationsstatistische Kennzahlen (Devi-

ance, LR- und CAIC-Statisik) zur Überprüfung der Datenpassung eines ein- oder mehrdimensionalen Modells herangezogen, um die reliable Erfassung und ausreichende Abgrenzung der ausgewiesenen Teilkompetenzen zu kontrollieren. Hierfür werden die Reliabilitäten, die Zusammenhänge der latenten Teilkompetenzen und die Deviance-Statistik betrachtet. Auf der Grundlage der empirischen Daten kann anhand der Modellgeltungstests entschieden werden, ob eine empirische Evidenz für die zugrundliegenden theoretischen Annahmen zur Kompetenz gegeben ist oder ob sie modifiziert werden müssen (vgl. Voss, 2009, S. 14 ff.).

2.3.3 STANDARDS FÜR DIE KOMPETENZORIENTIERTE LEISTUNGSMESSUNG

Für die kompetenzorientierte Leistungsmessung haben sich in der empirischen Bildungsforschung Standards herausgebildet, um Kompetenz im Sinne von „kontextspezifischen kognitiven Leistungsdispositionen“ zu erfassen. Diese Einengung auf die Kognition wird von der Fachdidaktik vielfach kritisiert (z. B. Hurrelmann, 2002). Sie ist jedoch forschungsmethodologisch bedingt, da die Leistungstests nur die kognitive Disposition erfassen können (vgl. Blatt, Müller & Voss, 2010). Weitere Aspekte, die Einfluss auf die Kompetenz haben können, z. B. Motivation oder Selbstkonzept, werden durch Befragungen erhoben.

Für eine kompetenzorientierte Leistungsmessung setzt die empirische Bildungsforschung folgende Standards:

- Testentwicklung auf der Grundlage eines fachspezifischen theoretischen Rahmenkonzepts.
- Den statistischen Gütekriterien Objektivität, Reliabilität und Validität entsprechenden Tests.
- Skalierung der Testergebnisse mithilfe von IRT-Verfahren und Modellgeltungstests zur Kompetenzmodellierung.
- Ermittlung von Einflussfaktoren auf die Kompetenz auf Basis der Befragungsdaten.
- Ermittlung von Kompetenzstrukturen und -profilen.

Eine den Standards der empirischen Bildungsforschung entsprechende kompetenzorientierte Leistungsmessung setzt also voraus, dass das Testinstrument im Einklang mit der domänenspezifischen Theorie entwickelt wird und den statistischen Gütekriterien genügt. Dazu werden die erfassten Leistungsdaten mit Verfahren der Item-Response-Theory (IRT) skaliert und die

Dimensionalität des angenommenen Kompetenzkonstrukts modelliert und durch Modellgeltungstests überprüft. Mit den Befragungsdaten können mögliche Einflussfaktoren auf die Kompetenz durch entsprechende multivariate Analyseverfahren, z. B. Strukturgleichungsmodelle oder Mehrebenenanalysen, ermittelt werden. Inwiefern sich Kompetenzstrukturen in den Daten ermitteln lassen bzw. Kompetenzprofile abgeleitet werden können, wird durch Clusteranalysen und Latent-Class-Analysen (LCA) geprüft.

3. STATISTISCHE ANALYSEMETHODE ZUR LÄNGSSCHNITTLICHEN KOMPETENZERFASSUNG

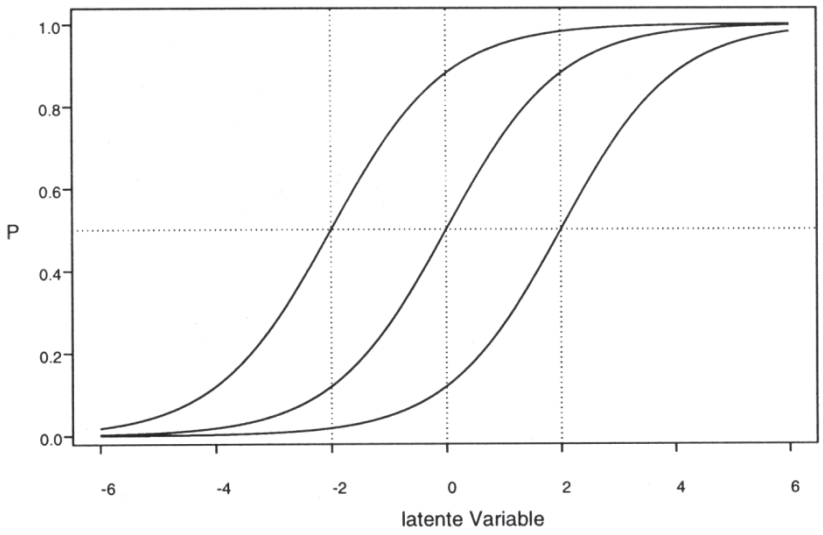
Zur längsschnittlichen Kompetenzerfassung können unterschiedliche statistische Analyseverfahren aus etablierten Schulleistungsstudien angewendet werden. Dazu zählen die Item-Response-Theory (IRT), die darauf basierenden Modelle der Veränderungsmessung und die längsschnittliche Kompetenzmodellierung sowie die Untersuchung von Einflussfaktoren mithilfe von Korrelations- und Regressionsanalysen.

3.1 ITEM-RESPONSE-THEORY

Die Item-Response-Theory (IRT) wird auch als probabilistisches Testverfahren bezeichnet und stellt das zentrale Verfahren bei der Kompetenzmodellierung dar (vgl. Bortz & Döring, 2006, S. 206). Sie grenzt sich von der klassischen Testtheorie (KTT) ab, da sie das Testergebnis lediglich als Indikator für eine latente Dimension ansieht (vgl. Moosbrugger, 2012, S. 232). Aufgrund der Abgrenzung von der KTT eignet sich das probabilistische Testverfahren besonders für die Modellierung von erziehungswissenschaftlichen Fragestellungen (vgl. Voss, 2006, S. 96). Ebenso zeigt die Entwicklung in der empirischen Bildungsforschung, dass sich die IRT für die Längsschnittanalyse bewährt hat (Rauch & Hartig, 2011, S. 253).

Konkret handelt es sich bei dem Verfahren um eine ein- oder mehrparametrige logistische Funktion, die auf Rasch (1960) zurückzuführen ist und zur Bestimmung einer Wahrscheinlichkeit zur Lösung eines Items in Abhängigkeit des sogenannten funktionalen Zusammenhangs zwischen der zugrundeliegenden Personenfähigkeit (Personenparameter) und Itemschwierigkeit (Itemparameter) dient. Die beiden Parameter können bei der IRT anhand der Itemcharakteristik (Item Characteristik Curve) auf einer gemeinsamen Skala abgebildet werden.

Abbildung 3.1: Itemcharakteristiken des Rasch-Modells



(Schnell et al. 2011, S. 189)

In Abbildung 3.1 sind drei Beispiele für eine Itemcharakteristik dargestellt, wobei die X-Achse die Personenfähigkeit und die Y-Achse die Lösungswahrscheinlichkeit wiedergibt (vgl. Bortz & Döring, 2006, S. 207). In diesem Fall handelt es sich um ein einparametriges Raschmodell, bei dem die Trennschärfe der Items auf den Wert 1 fixiert ist (vgl. Rost, 2004, S. 133). Demnach steigt mit zunehmender Kompetenzausprägung die Lösungswahrscheinlichkeit einer Aufgabe, weshalb z. B. bei einer mittleren Kompetenz die Aufgaben mit einer 50-prozentigen Lösungswahrscheinlichkeit richtig bearbeitet werden (vgl. Hartig & Jude, 2007, S. 34 f.).

Mithilfe der Formel des dichotomen logistischen Raschmodells zur Berechnung der Lösungswahrscheinlichkeit wird die latente Dimension der Personenfähigkeit (θ_v) in Bezug zur Itemschwierigkeit (σ_i) bestimmt:

Formel 3.1: Dichotomes logistisches Raschmodell

$$p(X_{vi} = 1) = \frac{\exp(\theta_v - \sigma_i)}{1 + \exp(\theta_v - \sigma_i)}$$

(Rost, 2004, S. 119)

Durch eine Transformation wird aus der Formel 3.1 eine lineare Funktion, die sich aus der Differenz von Personenfähigkeit und Itemschwierigkeit bildet. Demnach wächst mit steigender Fähigkeit die Wahrscheinlichkeit für die Lösung eines Items.

Formel 3.2: Logit-transformierte Lösungswahrscheinlichkeit

$$\log \frac{p(X_{vi}=1)}{p(X_{vi}=0)} = \theta_v - \sigma_i$$

(Rost, 2004, S. 118)

Die logit-transformierte Lösungswahrscheinlichkeit bewirkt eine lineare Abhängigkeit zwischen Item- und Personenparametern (vgl. Rost, 2004, S. 118). Anhand der Skalierungssoftware ConQuest 2.0 (vgl. Wu, Adams, Wilson & Haldane, 2007) wird die Itemschwierigkeit und Personenfähigkeit geschätzt.

3.1.1 ITEMPARAMETER

Allgemein wird vorausgesetzt, dass die manifesten Items eine Homogenität aufweisen und dieselbe latente Dimension messen (vgl. Moosbrugger, 2012, S. 229). Besonders wichtig ist hierbei die spezifische Objektivität, die besagt, dass die Schätzung unabhängig von der Item- und Personenauswahl immer zu demselben Resultat führt (vgl. Bortz & Döring, 2006, S. 209; Moosbrugger, 2012, S. 247). Ebenso ist die Eindimensionalität der Items von Bedeutung, die besagt, dass die Items unabhängig von ihrer Schwierigkeit das Testkonstrukt in kohärenter Weise messen (vgl. Rost, 2004, S. 370). Dies wird geprüft, indem der Item-Fit in Form des gewichteten *Mean Square Residual Fit (MNSQ)* betrachtet wird, dessen Erwartungswert bei 1 liegt (vgl. Wu, Adams, Wilson & Haldane, 2007). Demnach liegt bei einem Wert größer 1 eine schlechte Passung im Mittel und bei einem Wert kleiner 1 eine bessere Passung im Mittel vor. Die Entfernung von nicht modellkonformen Items erhöht den Item-Fit der restlichen Items. Darüber hinaus ist die *Diskrimination* bzw. *Trennschärfe* eines Items relevant. Ein Item mit einer schlechteren Passung im Mittel (>1) führt zu einer Verringerung der Diskrimination. Bei einer mittleren Passung von 1 diskriminieren die Items höher. Ein Item mit einer besseren Passung

im Mittel (<1) trägt zu einer hohen Diskrimination bei, wodurch besser zwischen leistungsschwachen und -starken Probanden differenziert wird.

Darüber hinaus wird durch eine Transformation des ungewichteten und gewichteten MNSQ-Wertes die *T-Statistik* eines Items angegeben, wodurch inferenzstatistisch überprüft wird, inwiefern ein Item modellkonform diskriminiert. Demnach ist die Diskrimination eines Items bis zum einem T-Wert von 1.96 modellkonform.

Ausgehend von den internationalen Schulleistungsstudien kann ein modellkonformes Intervall für den Item-Fit von 0.80 bis 1.20 (vgl. Adams, 2002, S. 105) und eine untere Grenze für die Diskrimination von 0.26 (vgl. OECD, 2005) angenommen werden, um die Qualität der empirischen Modellierung zu erhöhen. Dabei ist anzumerken, dass der Item-Fit gewissermaßen stichprobenabhängig ist und nicht verallgemeinert werden sollte, um eine Auswahl von modellkonformen Items vorzunehmen. Ebenso wird diskutiert, ob die Diskrimination eines Items ein geeignetes Verfahren zum Ausschluss eines Items darstellt (vgl. Embretson, 1996).

3.1.2 PERSONENPARAMETER

Für die Bestimmung des Personenparameters in Form des Logits-Werts stehen drei Arten von Schätzern zu Verfügung, um aus den beobachteten Indikatoren in Form der bearbeiteten Items die latente Kompetenz zu schätzen (vgl. Bortz, 2004; Kolen & Tong, 2010; Kolen, Tong & Brennan, 2011). Dazu zählen der ML-Schätzer „Maximum-Likelihood-Estimator" (MLE, Kendall & Stuart, 1973), der EAP-Schätzer „Expected-A-Posteriori-Estimator" (EAP, Bock & Mislevy, 1982) und der PV-Schätzer „Plausible Values" (PVs, Mislevy, Beaton, Kaplan & Sheehan, 1992):

– ML-Schätzer neigen durch den unvermeidlichen Messfehler dazu, die Varianz der Personenfähigkeit verzerrt zu schätzen (vgl. Hartig & Kühnbach, 2006, S. 30). Aus diesem Grund nimmt der gewichtete ML-Schätzer „Weighted-Likelihood-Estimator" (WLE, Warm, 1989) eine Korrektur der Varianz vor, indem die Personenfähigkeiten von Schülerinnen und Schülern, die sich am oberen oder unteren Kompetenzbereich befinden, verringert bzw. vergrößert werden. Infolgedessen können auch die Personenfähigkeiten von Schülerinnen und Schülern bestimmt werden, die keine bzw. alle Items richtig bearbeitet haben. Damit gilt der weiterhin messfehlerbehaftete WLE-Schätzer als bes-

ter Punktschätzer für die Bestimmung von individuellen Personenfähigkeiten der Schülerinnen und Schüler (vgl. Rost, 2004, S. 314 f.).

- Im Gegensatz dazu ist der sogenannte EAP-Schätzer trotzt Messfehler in der Lage, eine geringere Varianz der Personenfähigkeit zu schätzen (vgl. ebd., S. 316). Trotzdem stellt er keinen besseren Punktschätzer dar, weil es zu einer verzerrten Schätzung der Personenfähigkeit für besonders gute bzw. schlechte Schülerinnen und Schüler kommt (vgl. Frey, 2012, S. 285).
- Der PV-Schätzer zeichnet sich dagegen durch eine unverzerrte Bestimmung der Personenfähigkeit anhand von fünf Plausible Values aus, da er zur Schätzung neben den bearbeiteten Items auch ein sogenanntes Hintergrundmodell zur Modellierung der latenten Kompetenz heranzieht. Unter einem Hintergrundmodell sind beispielsweise Merkmale zu verstehen, die den sozioökonomischen Status (z. B. Geschlecht, Migration, Bildung) abbilden. So erfolgt die Schätzung der Personenparameter in Abhängigkeit des Hintergrundmodells, um Unterschiede bei der Personenfähigkeit zu lokalisieren (vgl. Hartig & Kühnbach, 2006, S. 30). Dadurch ist der PV-Schätzer zwar nicht zur optimalen Bestimmung von Personenfähigkeiten einzelner Schülerinnen und Schüler geeignet, aber er eignet sich für Large-Scale-Assessments mit Gruppenvergleichen (vgl. Scherer, 2012, S. 118).

Die Güte der Schätzung der Personenfähigkeit lässt sich mithilfe von Item- und Modellparametern bewerten, die nun näher vorgestellt werden.

3.1.3 MODELLPARAMETER

Zu den Modellparametern eines IRT-Modells gehören die Reliabilität, die latenten Korrelationen sowie die Deviance-Statistik, die nun näher erläutert werden.

Die Reliabilität (EAP/PV) eines Tests gilt als Hauptkriterium in der psychometrischen Testentwicklung. Darunter ist zu verstehen, wie genau das zu erfassende Persönlichkeitsmerkmal mit den Testaufgaben gemessen werden kann. Die Reliabilität eines Tests sollte über dem Wert von 0.70 liegen (vgl. Moosbrugger & Kelava, 2012, S. 11).

Die latenten Korrelationen erlauben Aussagen über die Kovariation der entsprechenden Testleistung auf der Teilkompetenzebene, indem sie quantifizierte latente Zusammenhänge unter-

schiedlicher Teilkompetenzen ausweisen. Dabei gilt, dass hohe Korrelationskoeffizienten zweier Teilkompetenzen auf redundante Informationen schließen lassen. In diesem Fall ist es nicht notwendig, die Teilkompetenzen voneinander zu trennen. Jedoch gilt gleichermaßen, dass Teilkompetenzen per se korreliert sind und einen Korrelationskoeffizienten über 0.75 aufweisen (vgl. Adams & Carstensen, 2002, S. 152-154).

Ein Kriterium zur empirischen Überprüfung eines Kompetenzmodells stellt der Deviance-Wert dar, der die Dimensionalität der Daten anhand unterschiedlich komplexer Kompetenzmodelle vergleicht und somit ein empirisches Maß zur Passung der theoretischen Modellstruktur und der empirischen Daten darstellt. Hierbei gilt, dass ein niedriger Deviance-Wert für eine höhere Erklärungskraft für die in den Daten enthaltenen Informationen steht (vgl. Voss, Carstensen & Bos, 2005). Dennoch ist dieses Maß nur unzureichend, da es weitere Modelleigenschaften vernachlässigt. Besser geeignet sind dagegen das Akaike's Information Criterion (AIC), Consistent AIC (CAIC) oder Bayes Information Criterion (BIC), die zusätzlich zur Deviance die Parameteranzahl und Stichprobengröße einbeziehen (vgl. Rost, 2004, S. 340).

3.2 MODELLE DER VERÄNDERUNGSMESSUNG

Die Erforschung von Veränderungen in Bezug auf Leistung und Kompetenz als sogenannte Veränderungsmaße über mehrere Messzeitpunkte bezieht sich auf globale Veränderungen in Bezug auf generelle Zuwächse einer ganzen Population (z. B. Geschlechter- bzw. Ländervergleich) oder personenspezifischer Veränderungen einzelner Schülerinnen und Schüler (vgl. Bortz, 2004, S. 272; Hartig & Kühnbach, 2006, S. 28).

Die Messung von Veränderungen zwischen zwei Messzeitpunkten birgt generelle methodische Probleme, da sie einem hohen Messfehler unterliegt und somit zu nicht reliablen Differenzen führt. Ebenso korreliert die Differenz aus zwei Messzeitpunkten negativ mit dem Kompetenzwert der ersten Messung, und es ist generell fraglich, ob die Tests zu den zwei Messzeitpunkten überhaupt in der Lage sind, identisch zu messen als Voraussetzung dafür, dass eine Differenz gebildet werden darf (vgl. Bortz, 2004, S. 273 ff.). Um dieser Problematik zu entgehen, kann der Differenzwert als abhängige Variable in Bedingungsanalysen verwendet werden, um die Wirkung von Einflussfaktoren korrekt zu schätzen. Eine andere Möglichkeit zur messfehlerfreien Bestimmung der Kompetenzveränderung stellt die Verwendung von Plausible Values dar (vgl. Hartig, Jude & Wagner, 2008, S. 38).

Eine Veränderung (*D*) wird dabei als Differenz zwischen zwei Personenfähigkeiten (θ) verstanden, wie es die nachfolgende Formel für zwei Messzeitpunkte (v) ausdrückt.

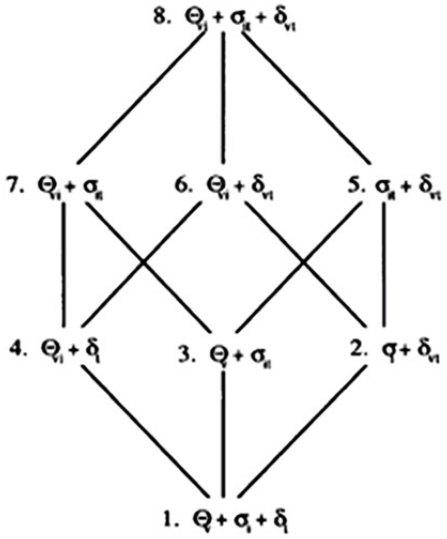
Formel 3.3: Differenzformel für Veränderungen

$$D_v = \theta_{v2} - \theta_{v1}$$

(Bortz, 2004, S. 273)

Für die Veränderungsmessung stehen IRT-Modelle zur längsschnittlichen Kompetenzmodellierung zur Verfügung (vgl. Meiser, 2007). Aufgrund der wiederholten Messung ist es nicht nur erforderlich, die getesteten Personen und verwendeten Items zu berücksichtigen, sondern ebenfalls eine Zuordnung zum jeweiligen Messzeitpunkt herzustellen. Für diesen Zweck bietet Rost (2004) eine dreifaktorielle Verallgemeinerung des Rasch-Modells (Rasch, 1960) mit insgesamt acht Modellen an, mit der jeweils die Lösungswahrscheinlichkeit für den Fall ausgeben wird, dass eine „Person (v) das Item (i) zum Zeitpunkt (t)" löst (vgl. Carstensen, Lankes & Steffensky, 2012, S. 110). Das System zur Verallgemeinerung des Rasch-Modells ist in Abbildung 3.2 veranschaulicht.

Abbildung 3.2: System der dreifaktoriellen Veränderungsmodelle



(Rost, 2004, S. 286)

Bei den acht Modellen zur Veränderungsmessung werden vier Stufen der Verallgemeinerung des Rasch-Modells unterschieden. Dabei repräsentiert das jeweils unterhalb liegende Modell einen Spezialfall des darüberliegenden Modells. Beide Modelle sind miteinander verbunden. Auf diese Weise sind sogenannte Modellgeltungskontrollen möglich, bei denen spezifische Modelle gegen allgemeine Modelle getestet werden (vgl. Bortz, 2004, S. 286). An dieser Stelle wird nur auf die zentralen Modelle eingegangen.

Die erste Stufe spezifiziert die globalen Veränderungen, die sich aus dem Zusammenspiel der Personenfähigkeit (θ_v), der Itemschwierigkeit (σ_i) sowie dem Einfluss des Messzeitpunktes (δ_t) für alle Personen und Items gleichermaßen ergeben. Folglich wird bei allen Personen und Items von einer identischen Lösungswahrscheinlichkeit ausgegangen, wie es der ersten Formel 3.4 für das dreifaktorielle Rasch-Modell zu entnehmen ist (vgl. ebd., S. 282).

Formel 3.4: Rasch-Modell für globale Veränderungen

$$p(X_{vit} = 1) = \frac{\exp(\theta_v + \sigma_i + \delta_t)}{1 + \exp(\theta_v + \sigma_i + \delta_t)}$$

(Bortz, 2004, S. 282)

Auf den weiteren Stufen werden spezifische Veränderungen erfasst, die mit jeweils unterschiedlichen Exponenten Anwendung in Formel 3.5 finden.

Formel 3.5: Rasch-Modell für spezifische Veränderungen

$$p(X_{vit} = 1) = \frac{\exp(\alpha_{vit})}{1 + \exp(\alpha_{vit})}$$

(Rost, 2004, S. 282)

Die zweite Stufe umfasst Modelle für spezifische Veränderungen „mit jeweils einer Wechselwirkung zwischen zwei Faktoren“ (Bortz, 2004, S. 284). Um personenspezifische Veränderungen abzubilden, müssen Wechselwirkungen zwischen Personen und Items zugelassen werden. Realisiert wird dies mit dem zweiten Modell, indem der doppelt indizierte Zeitpunkteffekt

(δ_{vt}) im Exponenten $\alpha_{vit} = \sigma_i + \delta_{vt}$ aufgenommen wird, um unterschiedliche Personenfähigkeiten zu den jeweiligen Messzeitpunkten zu erhalten. Sollen dagegen itemspezifische Veränderungen fokussiert werden, muss die doppelt indizierte Itemschwierigkeit (σ_{it}) berücksichtigt werden, wodurch sich der Exponent im dritten Modell zu $\alpha_{vit} = \theta_v + \sigma_{it}$ ändert.

Weitergehend umfasst die dritte Stufe der Verallgemeinerung zusätzliche Modelle „mit je zwei Wechselwirkungsparametern" (Bortz, 2004, S. 284). Ist beispielsweise eine Kombination aus personen- und itemspezifischen Veränderungen über die Messzeitpunkte erwünscht, so muss das Modell um den doppelt indizierten Itemparameter (σ_{it}) erweitert werden, wodurch sich der Exponent zu $\alpha_{vit} = \sigma_{it} + \delta_{vt}$ ändert, wie es das fünfte Modell zeigt.

Die vierte Stufe bildet folglich ein Modell mit „drei Wechselwirkungsparametern" ab, das allerdings in der Praxis nicht anwendbar ist und einen Spezialfall darstellt (vgl. Rost, 2004, S. 286).

Die Veränderung der Kompetenz bei einer längsschnittlichen Messung wird als Effekt bezeichnet, dessen Stärke als Effektgröße klassifiziert wird. Zur Ermittlung von Effekten wird in der Regel auf die Arbeit von Cohen (1988) zurückgegriffen.

Formel 3.6: Effektstärken nach Cohen

$$d = \frac{M_A - M_B}{\sigma}$$

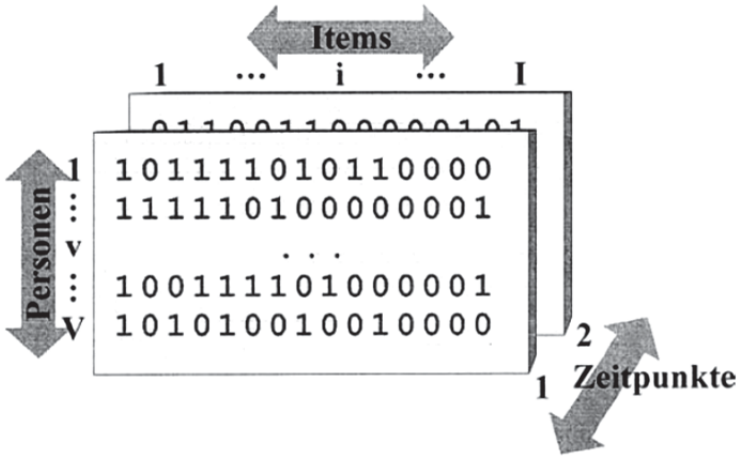
(Hartig et al., 2008; S. 48)

Der Effekt (d) wird dabei aus der mittleren Differenz ($M_A - M_B$) der zu untersuchenden Merkmale abhängig von der zugrundeliegenden Varianz (σ) ermittelt, wobei die Effektstärke nach kleinen (0.20), mittleren (0.50) und großen (0.80) Effekten unterteilt ist. Somit wird eine vergleichbare Maßeinheit geschaffen, die unabhängig von der ursprünglichen Skala einer Studie oder Analyse ist (vgl. Hartig et al., 2008; S. 47 f.).

3.3 VERFAHREN ZUR MODELLIERUNG VON LÄNGSSCHNITTLICHEN KOMPETENZENTWICKLUNGEN

Für die Kompetenzmodellierung im Längsschnitt haben sich Methoden der Item-Response-Theory (IRT) bewährt (vgl. Hartig & Kühnbach, 2006, S. 29). Dabei steht aus der Veränderungsmessung zu mehreren Messzeitpunkten eine Datenmatrix zur Verfügung, in der Personen zum jeweiligen Zeitpunkt eine Menge von Items bearbeitet haben. Grafisch ist dieser sogenannte dreidimensionale Datenkubus in Abbildung 3.3 wiedergegeben (vgl. Rost, 2004, S. 272).

Abbildung 3.3: Dreidimensionaler Datenkubus der Veränderungsmessung



(Hartig & Kühnbach, 2006, S.32)

Anhand von zwei Messzeitpunkten wird nun aufgezeigt, wie die längsschnittliche Kompetenzmodellierung im Fall von personenspezifischen Veränderungen erfolgt. Die Modellierung erfolgt auf der Grundlage zweier Testdesigns. Zum einen existieren identische Tests pro Messzeitpunkt, die über alle Items eine Verbindung zwischen den Erhebungen herstellen und auch als sogenannte Ankeritems bezeichnet werden (vgl. Andersen, 1985). Zum anderen gibt es unterschiedliche Tests pro Messzeitpunkt, die jeweils zeitpunktspezifische Items enthalten und über identische Ankeritems verbunden sind (vgl. Embretson, 1991). Bei beiden Testdesigns kommt die als „*Fixed Parameters Scale Linking*" bekannte Methode für die Ankeritems zum Einsatz, wobei die Itemschwierigkeit zum ersten Messzeitpunkt geschätzt und zur Fixie-

rung in den weiteren Messungen genutzt wird (vgl. von Davier, Carstensen & von Davier, 2008, S. 132). Diese Methode gilt als robust gegenüber Modellverletzung und einer daraus resultierenden verzerrten Schätzung der Item- und Personenparameter (vgl. Robitzsch 2009).

3.3.1 KOMPETENZMODELLIERUNG FÜR PERSONENSPEZIFISCHE VERÄNDERUNGEN

Um personenspezifische Veränderung der Kompetenz im Längsschnitt zu modellieren, ist es auf der Grundlage von identischen bzw. unterschiedlichen Tests möglich, die erfasste Kompetenz in getrennten oder gemeinsamen IRT-Modellen pro Messzeitpunkt zu schätzen. Es werden nun drei Verfahren zur längsschnittlichen Modellierung von personenspezifischen Veränderungen im Einzelnen aufgezeigt. Zu den drei Verfahren gehören die getrennte Skalierung (1), die Skalierung mit virtuellen Personen (2) und die Skalierung mit latenten Dimensionen (3). Die drei Verfahren werden exemplarisch anhand von identischen Tests vorgestellt. Im Anschluss wird am Beispiel der Skalierung mit virtuellen Personen das Verfahren für unterschiedliche Tests skizziert.

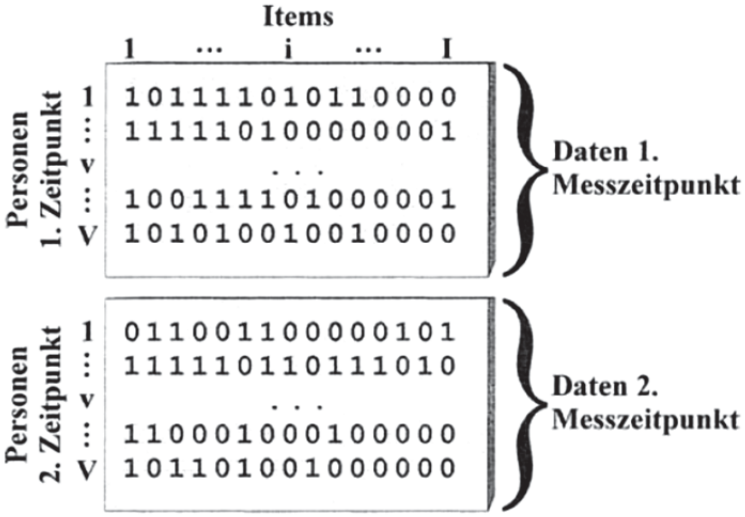
Bei der getrennten Skalierung (1) werden die Daten zur Schätzung der Item- und Personenparameter pro Messzeitpunkt als separate Modelle behandelt. In einem ersten Schritt ist es erforderlich, die Messinvarianz der beiden Messzeitpunkte zu prüfen, d. h., es werden die Itemparameter der beiden Messzeitpunkte frei geschätzt, um mittels eines sogenannten grafischen Modelltests mögliche Abweichungen zwischen zwei Teilstichproben der eigentlichen Stichprobe nach einem frei wählbaren Merkmal festzustellen. Sofern der Vergleich die Messinvarianz der Itemparameter bestätigt, wird im Sinne der Personenhomogenität für alle Personen eine homogene Fähigkeit bzw. ein Merkmal gemessen (vgl. Bühner, 2006, S. 343 f.). Es erfolgt eine weitere getrennte Skalierung des zweiten Messzeitpunktes, bei der allerdings die Itemparameter des ersten Messzeitpunktes zur Fixierung für den zweiten Messzeitpunkt genutzt werden, um die Veränderung der Personenfähigkeit zu ermitteln. Abschließend wird im Sinne der Vergleichbarkeit der Personenparameter eine Verlinkung zwischen den Messzeitpunkten vorgenommen, indem der ursprüngliche Mittelwert und die Standardabweichung zur Berechnung der weiteren Entwicklung hinzugezogen werden. Konkret wird von der geschätzten Personenfähigkeit zum zweiten Messzeitpunkt der Mittelwert des ersten Messzeitpunktes subtrahiert und durch die Standardabweichung aus der ersten Messung dividiert. Auf diese

Weise erhält man die Entwicklung der Personenfähigkeit in Abhängigkeit vom ersten Messzeitpunkt.

Zur Kompetenzmodellierung für personenspezifische Veränderungen mit einem gemeinsamen Skalierungsmodell für zwei Messzeitpunkte besteht die Möglichkeit, dies mithilfe von sogenannten virtuellen Personen oder einer weiteren latenten Dimension zur Abbildung der jeweiligen Messzeitpunkte umzusetzen.

Die Skalierung mit virtuellen Personen (2) reduziert die dreidimensionale Datenstruktur um die zeitliche Dimension (vgl. Abbildung 3.3) und fügt die Daten aus dem zweiten Messzeitpunkt als virtuelle Personen unterhalb des ersten Messzeitpunktes im Datensatz ein, wie es in Abbildung 3.4 dargestellt ist.

Abbildung 3.4: Kompetenzmodellierung mit virtuellen Personen



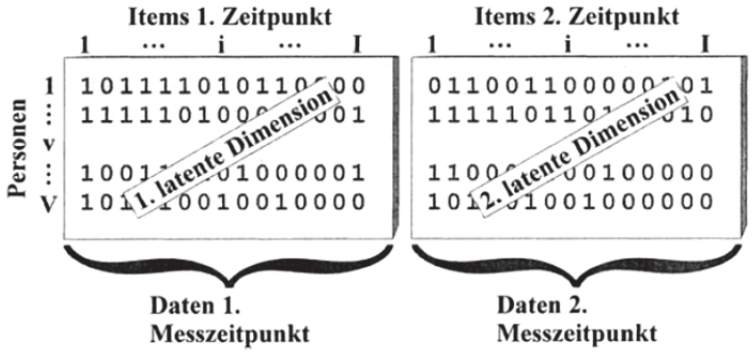
(Hartig & Kühnbach, 2006, S.32)

Dies bewirkt, dass für die virtuellen Personen angenommen wird, dass die getesteten Aufgaben von einer Stichprobe mit allen Schülerinnen und Schülern unabhängig vom Messzeitpunkt bearbeitet wurden. Demzufolge hat die Kompetenzmodellierung mit virtuellen Personen den Nachteil, dass bei der Skalierung keine Abhängigkeit zwischen den Messzeitpunkten besteht und folglich unabhängige Schätzungen der Personenfähigkeiten erfolgen, da nur die Itemschwierigkeiten automatisch fixiert sind. Mit einer Rückführung in die dreidimensionale

Datenstruktur kann dies korrigiert werden, indem anschließend für jede Schülerin und jeden Schüler die jeweiligen Personenfähigkeiten pro Messzeitpunkt feststehen (vgl. Hartig & Kühnbach, 2006, S. 33 f.).

Bei der Skalierung mit latenten Dimensionen (3) werden die jeweiligen Messzeitpunkte als latente Dimensionen modelliert, um personenspezifische Veränderungen zu schätzen. Der Aufbau des mehrdimensionalen Rasch-Modells (vgl. Rasch, 1961) mit zwei Messzeitpunkten ist der Abbildung 3.5 zu entnehmen.

Abbildung 3.5: Kompetenzmodellierung mit latenten Dimensionen



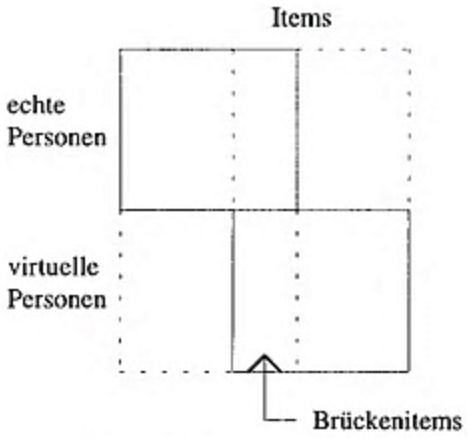
(Hartig & Kühnbach, 2006, S.35)

Danach sind die Lösungswahrscheinlichkeiten einer Aufgabe der jeweiligen latenten Dimensionen zu entnehmen, mit der die Schülerinnen und Schüler zu den unterschiedlichen Messzeitpunkten eine Aufgabe lösen. Aufgrund der Fixierung der Itemschwierigkeit können die Personenfähigkeiten pro Dimension auf einer gemeinsamen Skala verortet werden. Des Weiteren werden die Zusammenhänge zwischen den jeweiligen latenten Dimensionen bestimmt (vgl. Hartig & Kühnbach, 2006, S. 33 f.).

Alternativ ist es möglich, nur die Ankeritems zu fixieren, wenn sich die Anzahl der Items pro Messzeitpunkt durch zusätzliche zeitpunktspezifische Items unterscheidet. Dies ist beispielsweise der Fall, wenn im Verlauf der Schulzeit neue Lerninhalte berücksichtigt oder bestehende Items aus statistischen Gründen ausgeschlossen werden. Am Beispiel der Skalierung mit virtuellen Personen wird die Datenstruktur für diesen Fall veranschaulicht, wobei dasselbe Vorgehen auch für die zwei anderen Verfahren angewendet werden kann. So ist es möglich, die

personenspezifische Veränderung der Kompetenz über Ankeritems zu ermitteln, wie es in Abbildung 3.6 dargestellt ist.

Abbildung 3.6: Kompetenzmodellierung mit virtuellen Personen bei teilweiser Fixierung der Itemparameter



(Bortz, 2004, S. 281)

Bei diesem Verfahren werden nur die zu beiden Messzeitpunkten identischen Ankeritems einbezogen, während die übrigen Items nicht zur Schätzung der Veränderung herangezogen werden, da sie lediglich zum ersten oder zweiten Messzeitpunkt eingesetzt wurden (vgl. Rost, 2004, S. 281; Eggert, Bögeholz, Waltermann & Hasselhorn, 2010, S. 307).

Die drei vorgestellten Verfahren haben Vor- und Nachteile für die Modellierung von personenspezifischen Veränderungen, worauf in einem Vergleich kurz eingegangen wird. Die getrennte Skalierung (1) ermöglicht eine flexible Modellierung der Kompetenzen, indem sie in Bezug auf die Item- und Personenparameter anpassungsfähig ist, ohne jedoch einen direkten Bezug zwischen den Messzeitpunkten herzustellen. Aufgrund der automatischen Fixierung ist die Skalierung mit virtuellen Personen (2) kaum anpassungsfähig, stellt jedoch einen direkten Bezug zwischen den Messzeitpunkten her. Die Skalierung mit latenten Dimensionen (3) erweist sich dagegen als flexibles Verfahren zur Kompetenzmodellierung und setzt zudem die Messzeitpunkte in Beziehung zueinander. Inwiefern sich die empirische Umsetzung der drei Verfahren zur Modellierung von personenspezifischen Veränderungen realisieren lässt, wird in Kapitel 8.2 für die Modellierung der Rechtschreibkompetenz aufgezeigt.

3.4 KORRELATIONS- UND REGRESSIONSANALYSEN

Mithilfe von Korrelationsanalysen werden bivariate Beziehungen zwischen der zu untersuchenden Variable mit möglichen Einflussfaktoren aufgedeckt. Dabei erfolgt eine Analyse des Zusammenhangs zwischen den metrischen Variablen „Entwicklung der Kompetenz" bzw. „Veränderung der Kompetenzstruktur" und den meistens nominalen Merkmalen, die sich aus den „Einflussfaktoren" der Befragungsdaten über zwei Messzeitpunkte ergeben.

Infolge der Kombination von (dichotomen) nominalen und metrischen Variablen handelt es sich um eine punktbiseriale Korrelation, die mit der Statistiksoftware SPSS 22 nicht berechnet werden kann (vgl. Zöfel, 2002, S. 137). Daher muss zunächst der Korrelationskoeffizient nach Pearson (Produkt-Moment-Korrelation) ermittelt werden, der sich wiederum in die punktbiseriale Korrelation überführen lässt (vgl. Rasch, Friese, Hofmann & Naumann, 2006, S. 140).

Formel 3.7: Die Produkt-Moment-Korrelation

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \bar{y})}{(n-1) * s_x * s_y}$$

(Zöfel, 2002, S. 124).

Die Formel setzt sich aus der Kovarianz und der Standardabweichung der Variablen X und Y zusammen, wobei der Zusammenhang der Variablen mit einem Wert zwischen -1 und 1 ausgedrückt wird (vgl. Duller, 2006, S. 134).

Weitergehend lassen sich multivariate Beziehungen mit Regressionsanalysen auswerten. Dabei werden die Beziehungen zwischen einer abhängigen Variable (AV) und einer oder mehreren unabhängigen Variablen (UV) erklärt. Auf Basis der Auswertungsergebnisse kann eine Kausalbeziehung (Ursache-Wirkungs-Beziehung) zwischen der AV und UV hergestellt werden, die auch als „Je-desto-Beziehung" bezeichnet wird. Die nachfolgende Formel drückt die Beziehung zwischen AV und UV aus, indem die abhängige Variable durch die unabhängigen Variablen verändert wird (vgl. Backhaus, Erichson, Plinke & Weiber, 2008, S. 52 f.).

Formel 3.8: Kausalbeziehung zwischen AV und UV

$$Y = f(X_1, X_2, ..., X_{j,}..., X_J)$$

(Backhaus et al., 2008, S. 53)

Bei einer Vielzahl von unabhängigen Variablen handelt es sich um eine multiple Regression. Die multiple Regression basiert auf folgender Formel (vgl. Backhaus et al., 2008, S. 64):

Formel 3.9: Multiple Regressionsgleichung

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_j x_j + ... + b_J x_J$$

(Backhaus, et al., 2008, S. 64)

Die Schätzung der abhängigen Variable ($\hat{Y}$) wird durch die Konstante „b_0" sowie die Multiplikation des Regressionskoeffizientens „b_j" (j=1,2,…,J) mit dem Wert für die unabhängige Variable „X_j" (j=1,2,…,J) bestimmt (vgl. ebd., S. 58). Aufgrund verschiedener Skalenniveaus kann zusätzlich eine Standardisierung vorgenommen werden, um die dadurch entstehende Einwirkung auf die Regressionskoeffizienten zu umgehen (vgl. ebd., S. 66). Nicht nur die UV haben einen Einfluss auf die AV, sondern auch Messfehler können das Ergebnis beeinflussen. Deshalb gilt es, eine lineare Funktion zu finden, bei der die Abweichungen möglichst gering sind (vgl. ebd., S. 62). Unter Zuhilfenahme des korrigierten Bestimmtheitsmaß „R^{2}" wird abschließend überprüft, wie gut die abhängige Variable durch die unabhängigen Variablen erklärt werden kann. Im Gegensatz zu dem unkorrigierten Bestimmtheitsmaß werden hierbei die Anzahl der Untersuchungseinheiten, die unabhängigen Variablen sowie die Freiheitsgrade berücksichtigt, um einen zufällig entstandenen Erklärungsanteil zu korrigieren (vgl. ebd., S. 71 ff.).

Mit Korrelationsanalysen werden Zusammenhänge von Merkmalen ermittelt und mit Regressionsanalysen Effekte von unabhängigen Variablen auf eine abhängige Variable bestimmt.

4. EMPIRISCHE ERFORSCHUNG VON RECHTSCHREIBLEISTUNG UND -KOMPETENZ

Bei den national und international vergleichenden Leistungsstudien konzentriert sich die Kompetenzmessung auf Lesen, Mathematik und Naturwissenschaften (vgl. Kapitel 1.2.2). Obwohl die Rechtschreibung neben diesen Domänen gleichermaßen bedeutsam für die schulische und berufliche Entwicklung der Schülerinnen und Schüler ist, wurde die Erfassung der Rechtschreibleistung in der empirischen Bildungsforschung bisher nur in Form von Ergänzungsstudien berücksichtigt. Aus diesem Grund besteht Nachholbedarf bei der Erforschung der Rechtschreibkompetenz in nationalen und internationalen Vergleichsstudien zur Beschreibung und Erklärung des Erwerbs und der Entwicklung der Rechtschreibkompetenz (vgl. Schneider, Marx & Hasselhorn, 2008). Schneider, Marx & Hasselhorn (2008, S. 3) betonen, dass derzeit „eine zunehmende Nachfrage nach diagnostischen Verfahren zur Erfassung der Rechtschreibleistungen und nach Förderansätzen zur Verbesserung von Rechtschreibkompetenzen" bestehe. Die aktuelle Forschungsperspektive zur Diagnostik der Rechtschreibung wird durch das folgende Zitat aufgegriffen.

> „Rechtschreibleistung wird diagnostisch gesehen nicht mehr ausschließlich an der Menge von Rechtschreibfehlern festgemacht. Um herauszufinden, welche Stärken und Schwächen ein Schreiblerner hat, wird auf die Qualität der Schreibungen geschaut: Welcher Art sind die Fehler und welches Wortmaterial liegt vor? Diese Form der Diagnostik ist weniger ökonomisch, dabei aber sehr viel genauer und aussagekräftiger" (Fay, 2013, S. 11).

Die Rechtschreibleistung der Schülerinnen und Schüler wurde lange Zeit an der Anzahl der Rechtschreibfehler bemessen, wobei ein Wort als falsch gewertet wurde, unabhängig davon, ob ein oder mehrere Fehler darin enthalten waren und wie diese Fehler beschaffen waren. Auf dieser Grundlage können allerdings nicht die Schwächen und Stärken des Schreibers festgestellt werden, weshalb sich die Diagnostik der Rechtschreibung zunehmend auf die Fehlerart innerhalb der Testwörter bezieht und so die Rechtschreibleistung der Schülerinnen und Schüler differenziert erfasst. Böhme & Bremerich-Vos (2009) konkretisieren die Differenzierungsmöglichkeiten bei der Analyse von Fehlern und sprechen sich für eine qualitative Fehleranalyse aus:

„Bei der quantitativen Auswertung der Tests kann jedes falsch geschriebene Wort als ein Fehler angesehen werden. Es gibt aber viele verschiedene Möglichkeiten, ein Wort falsch zu schreiben, auch können pro Wort mehrere Fehler auftreten. Solche Differenzierungen sind für eine qualitative Fehleranalyse zentral, welche wiederum unabdingbar für (förder-) diagnostisch relevante Schlüsse ist. Die Fehlertypologie sollte orthografietheoretisch fundiert sein und nicht auf problematischen Kausalannahmen beruhen, was z. B. bei Kategorien wie »Flüchtigkeitsfehler« der Fall ist" (Böhme & Bremerich-Vos, 2009, S. 344 f.).

Hier stellt sich die grundlegende Frage nach einer angemessenen fachlichen Fundierung einer qualitativen Fehleranalyse. Sie kann nur mittels empirischer Untersuchungen verlässlich beantwortet werden. Dabei muss zwischen der Erfassung der Rechtschreibleistung und der Erfassung der Rechtschreibkompetenz unterschieden werden (vgl. Kapitel 2.1.3). Im Folgenden werden zunächst die in empirischen Bildungsstudien am Ende der Primarstufe und in der Sekundarstufe eingesetzten Tests zur Leistungsmessung der Rechtschreibung vorgestellt. Anschließend wird im Sinne einer kompetenzorientierten Leistungsmessung der Rechtschreibung gefragt, welche Tests und Studien dazu geeignet sind.

4.1 TEST- UND STUDIENÜBERBLICK

In diesem Überblick werden aus aktueller Perspektive Tests und Studien zur Erfassung der Rechtschreibleistung bzw. Rechtschreibkompetenz der Schülerinnen und Schüler zusammengestellt.

Die Darstellung der Testinstrumente zur Erfassung der Rechtschreibleistung in der Sekundarstufe erfolgt auf der Grundlage der Arbeiten von Frahm & Blatt (2011) sowie Frahm (2013). Dazu zählen die leistungsbezogenen Rechtschreibtests:

- Aachener förderdiagnostische Rechtschreibfehler-Analyse (AFRA)
- Analyse der Rechtschreibentwicklung (ANDREE)
- Deutscher Rechtschreibtest 3/4+ (DERET)
- Diagnostischer Rechtschreibtest 5 (DRT)
- Hamburger Schreibprobe 5-9 (HSP)

- Münsteraner Rechtschreibanalyse (MRA)
- Rechtschreibtest 6/7 (RST)
- Rechtschreibtest NRR (RST-NRR)
- Rechtschreibungstest (RT)
- Screening für Schul- und Bildungsberatung (SSB)
- Oldenburger Fehleranalyse (OLFA)

Aus den wenigen vorliegenden Leistungsvergleichsstudien, die sich mit der Domäne Rechtschreibung befassen, wurden die Studien LOGIK- und SCHOLASTIK als die ersten und größten Studien zur Entwicklung der Rechtschreibleistung in der Primarstufe ausgewählt. Zum anderen wird die Hamburger KESS-Studie zur längsschnittlichen Erfassung der Rechtschreibentwicklung im Kontext weiterer Kompetenzen von der Primarstufe bis zur Sekundarstufe I aufgegriffen.

Während sich zahlreiche Tests und Studien mit der Erforschung der Rechtschreibleistung befassen, liegen bislang zur kompetenzorientierten Leistungsmessung der Rechtschreibung lediglich drei Testinstrumente vor, die ausschließlich für den Einsatz in der Primarstufe konzipiert und in dazugehörigen Studien eingesetzt wurden. Dazu zählen der sprachsystematische Rechtschreibtest (Kapitel 5.2), der gutschrift-Test, dessen Anwendung in der gleichnamigen gutschrift-Studie zur empirischen Überprüfung des theoretischen Kompetenzmodells stattfand, sowie die Untersuchung von Böhme & Bremerich-Vos (2009) im Rahmen der zur Überprüfung der Bildungsstandards Vergleichsarbeiten (VERA) in Klassenstufe 3 und 4 (Kapitel 2.2.1) zur Erforschung der Rechtschreibkompetenz. Da der sprachsystematische Rechtschreibtest für diese Arbeit grundlegend ist, wird er in Kapitel 5 ausführlich behandelt.

4.1.1 TESTS UND STUDIEN ZUR LEISTUNGSMESSUNG DER RECHTSCHREIBUNG

In diesem Kapitel werden Tests, die für die Messung der Rechtschreibleistung bei Schülerinnen und Schülern in erster Linie für die Sekundarstufe konzipiert wurden sowie längsschnittliche Studien aus dem Primarstufenbereich vorgestellt.

TESTS ZUR LEISTUNGSMESSUNG DER RECHTSCHREIBUNG IN DER SEKUNDARSTUFE

Frahm & Blatt (2011) haben Testverfahren aus den letzten 20 Jahren für die Erfassung der Rechtschreibleistung in der Sekundarstufe gesichtet und die Bandbreite der Tests mithilfe von zehn ausgewählten Rechtschreibtests dargestellt.

Tabelle 4.1: Übersicht ausgewählter Rechtschreibtests

Kurztitel	Titel	Autor(en)
AFRA	Aachener förderdiagnostische Rechtschreibfehler-Analyse	Herné & Naumann (2002)
ANDREE	Analyse der Rechtschreibentwicklung	Köhler (2008)
DERET	Deutscher Rechtschreibtest 3/4+	Stock & Schneider (2008)
DRT	Diagnostischer Rechtschreibtest 5	Grund, Haug & Naumann (2004)
HSP	Hamburger Schreibprobe 5-9	May (2001)
MRA	Münsteraner Rechtschreibanalyse	Schönweiß (2004)
RST	Rechtschreibtest 6/7	Rieder (1992)
RST-NRR	Rechtschreibtest NRR	Bulheller, Ibrahimovic & Häcker(2005)
RT	Rechtschreibungstest	Kersting & Althoff (2004)
SSB	Screening für Schul- und Bildungsberatung	Kormann & Horn (2006)

(vgl. Frahm & Blatt, 2011, S. 547)

Eine detaillierte Beschreibung der zehn ausgewählten Rechtschreibtests ist bei Frahm & Blatt (2011) nachzulesen. Anhand der Kriterien führen die Autoren darüber hinaus eine vergleichende Analyse der Tests anhand der Kategorien Testformat, Art der Kompetenz, Zielgruppe, Zielsetzung und theoretische Grundlage durch. Zur Kategorisierung der Tests werden im Folgenden die Ergebnisse der vergleichenden Analyse dargestellt.

Die vergleichende Analyse der Rechtschreibtests nach den fünf Kategorien hat Folgendes ergeben: Das **Testformat** unterscheidet sich bezüglich der Testteile, wobei eine Konzentration bei den Lückendiktaten (AFRA, DERET, DRT, MRA, RST, RST-NRR, RT, SSB) neben den Fließtextdiktaten (AFRA, DERET), Wortdiktaten (ANDRE, HSP, SSB), Satztextdiktaten (HSP), Wortlisten (AFRA, SSB) und dem Korrekturlesen (HSP, RST) festgestellt werden konnte. Das Testfor-

mat entscheidet gleichzeitig über die enthaltenen Rechtschreibinhalte, d. h., dass beispielsweise die Getrennt- und Zusammenschreibung nicht verlässlich mit Wortlisten oder Lückendiktaten überprüft werden kann. Bei der **Art der Kompetenz** kann zwischen produktiven (Anfertigung eines Diktats) und reflexiven (Korrekturlesen) Leistungen unterschieden werden. Die zusammengestellten Rechtschreibtests sind für mehrere **Zielgruppen** konzipiert. So sind Tests sowohl für die Primar- als auch Sekundarstufe (AFRA, ANDREE, DERET, SSB), nur für die Sekundarstufe (DRT, HSP; MRA, RST) oder bestimmte Altersgruppen (RST-NRR, RT) geeignet. Hinsichtlich der **Zielsetzung** der ausgewählten Tests konnte festgestellt werden, dass die Tests der Förderdiagnose im unteren Leistungsbereich (DERET, DRT, SSB) der Diagnose der Rechtschreibleistung aller Leistungsgruppen (AFRA, ANDREE, HSP, MRA) sowie der Feststellung der allgemeinen Rechtschreibleistung (RST, RT) dienen. Dabei erfolgt eine qualitative Analyse der Rechtschreibleistung, mit der – im Gegensatz zur überwiegend quantitativen Rückmeldung auf Fehlerbasis – differenzierte Aussagen möglich sind. Die **theoretische Grundlage** lässt sich bei den Rechtschreibtests nur ansatzweise oder gar nicht feststellen, da nur wenige Auswertungskriterien für die Tests (AFRA, ANDREE, DERET, DRT, HSP, MRA) theoretisch begründet sind. Die Auswertungskriterien decken Anforderungen an die Lernenden und orthografische Regeln ab. Insgesamt wurde festgestellt, dass die Tests den vorgesehenen Rechtschreibinhalten der Rahmen- und Bildungspläne für die Sekundarstufe nicht voll entsprechen. Die vergleichende Analyse kommt zu dem Fazit, dass die vorliegenden Tests nur bedingt zur Diagnose geeignet sind. Die ausgewählten Tests sind zudem nicht zur kompetenzorientierten Leistungsmessung geeignet, da als Testgrundlage ein empirisch geprüftes Kompetenzmodell fehlt.

Ergänzend zu den in Frahm & Blatt (2011) analysierten zehn Rechtschreibtests, analysiert Frahm (2013) die Oldenburger Fehleranalyse (OLFA), für die eine empirische Überprüfung der fachlichen Fundierung durchgeführt wurde (vgl. Thomé & Gomolka, 2007).

Die Oldenburger Fehleranalyse ist testunabhängig und daher darauf ausgelegt, aus freien Texten bzw. Diktaten mit mindestens 350 Wörtern und 60 Fehlern für die Jahrgangsstufen 3 bis 9 die orthografische Kompetenz und Leistung zu ermitteln und Fördermaßnahmen aus den Ergebnissen abzuleiten. Das der OLFA zugrundeliegende Verständnis der Begriffe Kompetenz und Leistung gibt das folgende Zitat wieder.

„Die **orthographische Kompetenz** soll in unserem Zusammenhang definiert werden als ein Bündel von Fähigkeiten, das den schriftsprachlichen Produktionen notwendig zugrunde liegt. Demgegenüber wird die **orthographische Leistung**, die sich unmittelbar auf der Oberfläche (oder in den Produkten) zeigt, vorrangig durch die absolute Fehlerzahl oder etwas differenzierter mit den Fehlerzahlen in den einzelnen Rechtschreibbereichen definiert." (Thomé & Thomé, 2010, S. 11).

Mit dem Ziel, ein individuelles Monitoring der Rechtschreibleistung zu realisieren, kann das Verfahren flexibel eingesetzt werden. Für die Analyse steht eine Fehlerliste mit 37 Kategorien zur Verfügung, die sich an entwicklungspsychologischen Theorien orientiert und in Anlehnung an Frith (1985) analog zur HSP ein dreiphasiges Entwicklungsmodell (vor- bis protoalphabetisch, alphabetisch, orthografisch) annimmt. Auf der Grundlage der Fehlertypen werden Kompetenz- und Leistungswerte berechnet. Mit dem Kompetenzwert soll das latente grundlegende Rechtschreibwissen abgebildet werden, indem der prozentuale Anteil der alphabetischen und orthografischen Strategie addiert und der prozentuale Anteil der protoalphabetischen Strategie subtrahiert wird (vgl. Thomé & Thomé, 2010, S. 32). Dieser Leistungswert wird als Maß für das manifeste Können herangezogen, wobei die Gesamtzahl der Fehler auf 100 Wörter ermittelt und mithilfe von Richtwerten eingeordnet wird. Die Analyse der Fehler soll als Grundlage für eine anknüpfende individuelle Förderung dienen (vgl. ebd.).

Eine allerdings nur ansatzweise empirische Überprüfung der Auswertungskategorien wurde für die 9. Jahrgangsstufe mit Daten der DESI-Studie mittels der Item-Response-Theory (IRT) durchgeführt. Nach den Autoren konnte „die Zuordnung der Fehlerkategorien zu unterschiedlichen Kompetenzniveaus, die in der OLFA-Liste den Gruppen I bis III entsprechen, […] mithilfe der Rasch Skalierung aus den Daten der DESI-Studie bestätigt werden" (Thomé & Thomé, 2010, S. 13). Dies ist allerdings nicht nachprüfbar, da die Autoren keine konkreten Ergebnisse liefern.[4]

Zieht man ein Fazit zu den vorliegenden Testinstrumenten zur Erforschung der Rechtschreibleistung, so zeigt sich, dass für die unterschiedenen Fehlertypen eine empirische Fundierung weitgehend aussteht (vgl. Böhme & Bremerich-Vos, 2009).

[4] Daher werden Test und Studie in dieser Arbeit nicht zur kompetenzorientierten Leistungsmessung gezählt.

STUDIEN ZUR LEISTUNGSMESSUNG DER RECHTSCHREIBUNG

Im Folgenden werden die drei Studien LOGIK, SCHOLASTIK und KESS zur Erfassung der Rechtschreibleistung referiert, die sich durch ein längsschnittliches Testdesign auszeichnen. Eine abschließende Beurteilung der Studien erfolgt unter Zuhilfenahme der in Kapitel 2.4.3 abgeleiteten Standards für empirische Längsschnittuntersuchungen.

LOGIK- UND SCHOLASTIK-STUDIE

Die beiden von 1984 bzw. 1987 bis 2004 durchgeführten Münchener Studien LOGIK (Longitudinalstudie zur Genese individueller Kompetenzen) und SCHOLASTIK (Schulorganisierte Lernangebote und Sozialisation von Talenten, Interessen und Kompetenzen) verfolgten das Ziel, die Kompetenzentwicklung von Schülerinnen und Schülern darzustellen und unterschiedliche Entwicklungsverläufe zu erklären. In der Primarstufe fanden die beiden Studien von der ersten bis zur vierten Jahrgangsstufe parallel statt. Die Stichprobe konzentrierte sich auf Grundschulen aus dem ländlichen und städtischen Raum Münchens. Die multiperspektivisch angelegten Studien beziehen neben den Testpersonen (LOGIK N~200; SCHOLASTIK N~1200) auch deren Mitschülerinnen und -schüler, Lehrkräfte sowie Eltern in die Untersuchung ein. Dieses Studiendesign verfolgt das Ziel, individuelle Entwicklungsverläufe der Kompetenzen in Abhängigkeit von Einflüssen im Klassenkontext und von sozialen Merkmalen zu erfassen und zu erklären. Es fehlt allerdings eine Definition der Begriffe Rechtschreibleistung und -kompetenz. Die Rechtschreibleistung der Schülerinnen und Schüler wurde mithilfe eigens entwickelter Rechtschreibproben längsschnittlich erhoben. Bis zur 2. Jahrgangsstufe wurden Lückentexte und ab der 3. Jahrgangstufe ganze Sätze zur Messung der Rechtschreibleistung eingesetzt, die eine Mischung von Wörtern aus dem Grundwortschatz und unbekannten Wörtern beinhalten und der jeweiligen Jahrgangsstufe bezüglich der Schwierigkeit der Wörter angepasst sind. Die Daten wurden deskriptiv auf Ebene des ganzen Wortes anhand der prozentualen Lösungshäufigkeit ausgewertet und die Ergebnisse als Mittelwerte und Standardabweichungen sowie der minimalen bzw. maximalen Anzahl richtiger Schreibungen berichtet (vgl. Schneider, 2008; Weinert & Helmke, 1997).

KESS-STUDIEN

Bei den KESS-Studien (Kompetenzen und Einstellungen von Schülerinnen und Schülern) handelt es sich jeweils um eine Vollerhebung der Hamburger Schülerschaft mit ca. 14.000 Schülerinnen und Schülern in der vierten, siebten, achten, zehnten und elften Jahrgangstufe (KESS 4, 7, 8, 10/11) (vgl. Bos, Bonsen & Gröhlich, 2009; Bos & Gröhlich, 2010; Bos & Pietsch, 2006; Vieluf, Ivanov & Nikolova, 2011). In den KESS-Studien wurden unterschiedliche Rechtschreibtests zur Messung der Rechtschreibleistung eingesetzt. In der KESS 4 Studie aus dem Jahr 2003 wurden die zwei standardisierten Rechtschreibtests Hamburger Schreibprobe (HSP) und Dortmunder Schriftkompetenzentwicklung (DoSE) eingesetzt (Bos & Pietsch, 2006). Die KESS-Studien 7, 8 und 10 aus den Jahren 2003, 2005 und 2009 verwenden eigens für KESS entwickelte Lückentests zur Erfassung der Rechtschreibleistung in mehreren Testvarianten (KESS 7: 2 Testvarianten; KESS 8 & 10: 19 Testvarianten). Den zur Leistungsmessung eingesetzten Rechtschreibtests HSP und DoSE[5] liegen fachwissenschaftlich begründete Annahmen zur Rechtschreibkompetenz zugrunde, für die jedoch eine empirische Validierung aussteht. Die Tests beziehen sich auf eine Auswahl von 24 Ankeritems aus den beiden KESS 4 Tests. Zudem kommen im Verlauf der Sekundarstufe neue Wörter hinzu, wodurch in KESS 7 insgesamt 50 Testwörter und in KESS 8 & 10 insgesamt 60 Testwörter zur Verfügung stehen. In KESS 11 wurde die Rechtschreibleistung der Schülerinnen und Schüler nicht erfasst (vgl. Bos, Bonsen & Gröhlich, 2009; Bos & Gröhlich, 2010; Vieluf, Ivanov & Nikolova, 2011). Die längsschnittliche Auswertung der Rechtschreibleistungsdaten auf Basis der Wortschreibungen wird in den KESS-Studien mittels der Item-Response-Theory zur Bildung einer sogenannten Orthografie-Skala als Maß für den Lernstand in der Rechtschreibung durchgeführt, die auf einen Mittelwert von 100 und einer Standardabweichung von 30 normiert wird. Auf dieser Grundlage werden die Ergebnisse anhand von Mittelwerten und Standardabweichungen berichtet. Darüber hinaus wurde die Entwicklung durch Effektstärken nach Cohen (1988) beurteilt.

Eine Analyse der drei ausgewählten Längsschnittuntersuchungen nach den in Kapitel 1.2.3 zusammengestellten Standards ergibt im Hinblick auf die Stichprobenziehung, dass diese nach inhaltlichen Aspekten im Hinblick auf die Zielsetzung der Studie erfolgte. Die drei Studien zielen auf die Erforschung von Leistungsentwicklungen von Schülerinnen und Schülern ab. Die in KESS 7 bis KESS 10 eingesetzten Testinstrumente sind dem Lernfortschritt angepasst und

[5] Es existiert der gutschrift-Test als Weiterentwicklung von DoSE (vgl. Kapitel 4.1.2).

ermöglichen eine Verlinkung mit den Ergebnissen von KESS 4. In der LOGIK- bzw. SCHOLASTIK-Studie werden die Rechtschreibproben in der Schwierigkeit der jeweiligen Jahrgangsstufe angepasst und mit Ankeritems verbunden. Zur Auswertung wird eine Vielzahl von statistischen Methoden eingesetzt, die aus der deskriptiven und induktiven Statistik sowie den probabilistischen Testmodellen stammen. Die Ergebnisse werden zur Rückmeldung von Mittelwerten und Standardabweichungen auf der Ebene der prozentualen Lösungshäufigkeit sowie Effektstärken nach Cohen genutzt. Ebenso wird bei KESS die Rechtschreibkompetenz auf einer Kompetenzskala zurückgemeldet, obwohl kein empirisch validiertes Kompetenzmodell vorliegt.

Zusammenfassend lässt sich sagen, dass die KESS-Studie den Standards für längsschnittliche Untersuchungen entspricht und somit die standardisierte Erforschung der Rechtschreibleistung gewährleistet. Hingegen weisen die LOGIK- und SCHOLASTIK-Studie diesbezüglich Schwächen auf, die vermutlich der Pionierarbeit auf diesem Gebiet geschuldet sind. Allerdings fehlt es allen Studien an der theoretischen Grundlage eines Kompetenzmodells, weswegen die Erfassung der Rechtschreibkompetenz im bildungswissenschaftlichen Sinn mit diesen Studien nicht möglich ist und korrekterweise von der Erfassung der Rechtschreibleistung gesprochen werden muss.

4.1.2 TESTS UND STUDIEN ZUR KOMPETENZORIENTIERTEN LEISTUNGSMESSUNG DER RECHTSCHREIBUNG

Mit dem Ziel einer kompetenzorientierten Leistungsmessung der Rechtschreibung wurden bislang neben dem sprachsystematischen Rechtschreibtest (Kapitel 5.2) die Tests gutschrift (Löffler & Meyer-Schepers, 2001) und ein VERA-3-Test (Böhme & Bremerich-Vos, 2009) auf der Grundlage der Aachener Förderdiagnostischen Rechtschreibfehler-Analyse (AFRA, vgl. Kapitel 4.1.1) in Leistungsstudien in der Primarstufe eingesetzt.

Mit dem gutschrift-Test wurde im Rahmen der gutschrift-Studie im Jahr 2006 die Rechtschreibkompetenz von Zweitklässlern erhoben. Der gutschrift-Test stellt eine Weiterentwicklung des sogenannten DoSE-Tests (Dortmunder Schriftkompetenzentwicklung) dar. Anhand der gutschrift-Studie erfolgte die empirische Überprüfung des theoretischen Kompetenzmodells des gutschrift-Tests (vgl. Voss, Löffler, Meyer-Schepers, Meckel & Kowalski, 2008). Aus diesem Grund wird einleitend die Studie skizziert, um nachfolgend den Test darzustellen.

GUTSCHRIFT

An der gutschrift-Studie nahmen 29 ausgewählte Dortmunder Grundschulen mit insgesamt 1.578 Schülerinnen und Schülern der zweiten Klassenstufe im Schuljahr 2006/07 teil (vgl. Voss, et al., 2008). Die nicht zufällig gezogene Stichprobe wurde in eine Treatment- und Vergleichsgruppe eingeteilt, wobei diesbezüglich die Kriterien in den vorliegenden Veröffentlichungen nicht weiter spezifiziert werden. Im Vordergrund stand die Frühdiagnose rechtschreibschwächerer Schülerinnen und Schüler auf der Grundlage eines differenziellen theoretischen Rahmenkonzepts, das elementare und erweiterte Fähigkeiten ausweist. Gleichzeitig wurden unterschiedliche Unterrichtskonzepte bzw. Fortbildungsmaßnahmen in der Treatmentgruppe verglichen. Dazu wurden an zwei Messzeitpunkten die elementaren und erweiterten Fähigkeiten mit demselben gutschrift-Test gemessen.

Der gutschrift-Test umfasst einen Lückentext mit 35 Testwörtern für die Jahrgangsstufen 1 bis 4. Der Test wird mithilfe einer quantitativen Analyse der richtiggeschriebenen Wörter und einer computergestützten qualitativen Fehleranalyse ausgewertet. Die Fehleranalyse basiert auf einem theoretischen Rahmenkonzept, wobei zwei Teilkompetenzen (phonographische und grammatische Teilkompetenz) in vier Kompetenzstufen (elementar-phonographisch und -grammatisch sowie erweitert-phonographisch und -grammatisch) unterschieden werden (vgl. Löffler & Meyer-Schepers, 2001). Die Datenauswertung der 35 Testwörter erfolgte auf Grundlage der deskriptiven Statistik und der probabilistischen Testtheorie, die zur Rückmeldung von Mittelwerten auf der Ebene der prozentualen Lösungshäufigkeit sowie der Standardabweichungen auf Gruppen- und Individualebene genutzt wurden. Ebenso wurden die Kompetenzentwicklung durch Effektstärken und die ein- und mehrdimensionale Datenmodellierung anhand des Deviance-Wertes (vgl. Wu et al., 2007) beurteilt. Die empirische Modellierung des theoretischen Rahmenkonzepts erfolgte im Rahmen der gutschrift-Studie, die das zweidimensionale Kompetenzmodell mit Reliabilitäten von 0.63 bzw. 0.76 für die phonografische und grammatische Teilkompetenz sowie einen korrelativen Zusammenhang in Höhe von 0.58 ausweist (vgl. Voss et al., 2008). Daraus schließen die Autoren auf eine empirische Evidenz des Kompetenzmodells, das elementare und erweiterte phonographische und grammatische Fähigkeiten ausdifferenziert. Die Befundlage ist jedoch nach Kriterien der internationalen Schulleistungsstudien als grenzwertig zu bezeichnen.

VERA-3-TEST AUF DER GRUNDLAGE DES AFRA-TESTS

Im Rahmen der Testpilotierung für die Überprüfung der Bildungsstandards in der Rechtschreibung in VERA-3 (vgl. Kapitel 2.2.1) wurde im Frühjahr 2006 ein an den AFRA-Test (vgl. Kapitel 4.1) angelehnter Rechtschreibtest parallel in der 3. und 4. Jahrgangsstufe eingesetzt (vgl. Böhme & Bremerich-Vos, 2009; Blatt & Frahm 2013). Damit sollte auf quantitativer Ebene das globale Kompetenzniveau der Schülerinnen und Schüler ermittelt werden und auf qualitativer Ebene eine Fehleranalyse im Hinblick auf die orthografischen Stufen des Rechtschreiberwerbs durchgeführt werden (vgl. Herné & Naumann, 2005).

Als Testinhalt wurden aus den KMK-Bildungsstandards für das Fach Deutsch in der Primarstufe die ersten beiden Kompetenzaspekte ausgewählt, also die normgerechte Schreibung rechtschreibwichtiger Wörter, die Anwendung des Einprägens sowie Ableitens als zwei von drei genannten Rechtschreibstrategien.

Für die Untersuchung standen insgesamt 80 als didaktisch sinnvoll zu bezeichnende Testwörter aus dem Orientierungswortschatz (Naumann, 1999) zur Verfügung, die z. B. problematische Phonem-Graphem-Korrespondenzen enthalten. Jeweils zwei Testwörter wurden in insgesamt 40 Lückensätze eingesetzt. Daraus wurden vier Testvarianten mit jeweils 20 Testwörtern gebildet. Drei Varianten wurden in beiden Jahrgangsstufen und der weitere Test nur in der 4. Jahrgangsstufe eingesetzt. Für die Bearbeitung der Tests waren 20 Minuten veranschlagt, in denen ein Testleiter sowohl die Instruktionen zu den Tests als auch den eigentlichen Lückentext vorgelesen hat.

Die Daten von 3480 Schülerinnen und Schülern aus der 3. und 4. Jahrgangsstufe wurden für die quantitative und qualitative Analyse sowie die empirische Überprüfung des Kompetenzmodells genutzt. Die quantitative Auswertung erfolgte aufgrund der Richtigschreibung der Testwörter und die qualitative Auswertung anhand sogenannter Lupenstellen bzw. Fehlerkategorien. Die Analyse ergab, dass der Test geeignet ist, die globale Rechtschreibleistung mittels der Testwörter für beide Jahrgangsstufen reliabel zu erfassen. Ebenso wurden die Fehler mehrheitlich durch die neun definierten Fehlerkategorien abgedeckt. Es gelang jedoch unter Anwendung unterschiedlicher statistischer Verfahren (Hauptkomponentenanalyse, Rasch-Skalierung, kognitive Diagnosemodelle) nicht, die neundimensionale Struktur des theoretischen Kompetenzkonstrukts nachzuweisen. Es deutete sich vielmehr eine eindimensionale Modellstruktur an.

Demnach lässt sich festhalten, dass die beiden Testinstrumente geeignet sind, die Rechtschreibleistung in der Primarstufe zu erfassen, im Hinblick auf eine kompetenzorientierte Leistungsmessung jedoch Einschränkungen vorliegen. Dies könnte auf eine eingeschränkte Eignung der zugrundeliegenden fachlichen Rahmenkonzepte hinweisen (vgl. Blatt & Frahm, 2013, S. 16 ff.).

5. SPRACHSYSTEMATISCHES RECHTSCHREIBKOMPETENZ-MODELL

In diesem Kapitel wird das sprachsystematische Rechtschreibkompetenzmodell zusammenfassend dargestellt (vgl. auch im Folg. Blatt et al., 2011; Voss, Blatt & Kowalski, 2007). Zunächst werden die sprachwissenschaftlichen und didaktischen Grundlagen aufgezeigt, um anschließend das theoretische Rahmenkonzept und den sprachsystematischen Rechtschreibtest zu beschreiben. Abschließend werden die Kompetenzmodellierung im Kontext der entsprechenden Studien und die Ergebnisse dargelegt.

5.1 SPRACHWISSENSCHAFTLICHE UND DIDAKTISCHE GRUNDLAGEN

In der Rechtschreibdidaktik hat sich Ende der 1990er-Jahre im Anschluss an die Grundlagenforschung von Hinney (1997) ein Paradigmenwechsel vollzogen. Während sich die Lerninhalte der Rechtschreibung bis dahin am amtlichen Regelwerk ausrichteten, erschloss Hinney die Ergebnisse der Graphematik für die Rechtschreibdidaktik. Die Graphematik ist ein Teilgebiet der Grammatik und beschreibt, wie man schreibt, während das amtliche Regelwerk festlegt, wie richtig geschrieben wird (vgl. Fuhrhop, 2006, S. 1). Hinney (1997) stellte die These auf, dass die im amtlichen Regelwerk festgelegten Regeln für Rechtschreibkönner, aber nicht für Lernende hilfreich seien, und nahm eine Neubestimmung der Rechtschreibinhalte vor, indem sie die graphematischen Forschungsergebnisse aufgriff. Diese weisen die deutsche Wortschreibung als in ihrem Kernbereich sehr systematisch geregelt aus, während das auf Konrad Duden zurückgehende amtliche Regelwerk die deutsche Rechtschreibung als ein kompliziertes Regelwerk mit zahllosen Ausnahmen darstellt. Dies hängt damit zusammen, dass die Graphematik das Schriftsystem als eigenständiges System begreift und untersucht und erst in einem zweiten Schritt mit dem Lautsystem in Beziehung setzt (Interdependenzthese), während Konrad Duden die Schrift als ein Abbild der gesprochenen Sprache betrachtete (Dependenzthese).

Die hohe Systematik der graphematisch bestimmten Lerninhalte eignet sich in besonderer Weise für ein theoretisches Rahmenkonzept zur empirischen Kompetenzmodellierung. Darüber hinaus eröffnen die graphematischen Forschungsergebnisse einen erweiterten Blick auf den Lerngegenstand Rechtschreibung. Darin liegen Chancen für einen effizienteren Recht-

schreibunterricht, was angesichts vorliegender Forschungsergebnisse mehr als wünschenswert erscheint. So konstatieren Valtin, Badel, Löffler, Meyer-Schepers, & Voss (2003), dass „unter Lernzielaspekten der ermittelte Leistungsstand in der Rechtschreibung nicht befriedigen [kann] – schon gar nicht, wenn man die für den Rechtschreibunterricht aufgewendete Zeit berücksichtigt." (S. 257). Es konnte ein Mangel an grundlegenden Einsichten der Schülerinnen und Schüler in die Struktur der Wortschreibung nachgewiesen werden (vgl. ebd., S. 257).

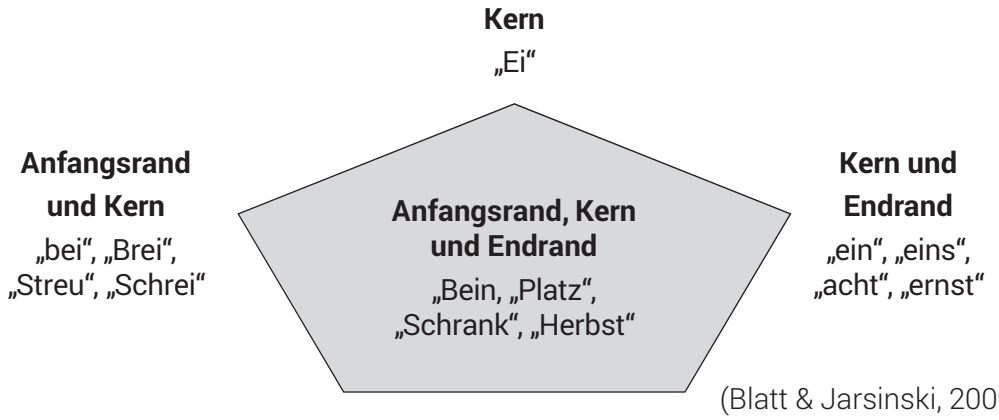
Im Folgenden werden zunächst die graphematischen Erkenntnisse und ihre didaktische Relevanz zusammengefasst, um im Anschluss daran das theoretische Rahmenkonzept für die Kompetenzmodellierung darzustellen.

Die graphematische Forschung zeigt auf, dass der historische Sinn der Rechtschreibung in der Erleichterung der Texterfassung für den Leser liegt (vgl. auch im Folg. Eisenberg & Fuhrhop, 2007). Weiterhin wird das Schriftsystem nicht wie im amtlichen Regelwerk als ein zur gesprochenen Sprache sekundäres System betrachtet, sondern als ein eigenständiges, das Teil des größeren Sprachsystems ist. Daher stellt Rechtschreiblernen keinen abgekoppelten Bereich des allgemeinen sprachlichen Lernens dar, sondern ist ein zentraler Bestandteil des sprachlichen Lernens (vgl. Blatt, 2010, S. 101 f.). Im Prozess des Rechtschreiblernens wird ein hohes Lernpotenzial zur Entwicklung der mündlichen und schriftlichen Sprachkompetenz gesehen (Eisenberg, 2004; Eisenberg & Fuhrhop, 2007). Damit wird der Rechtschreibung eine besondere und wichtige Rolle zugesprochen, die auch Schülerinnen und Schülern einsichtig gemacht werden kann, um ihre Lernmotivation zu erhöhen.

Die kognitive Durchdringung des Schriftsystems kann durch den Nachweis der graphematischen Forschung, dass die deutsche Rechtschreibung im Kernbereich systematisch durch nur vier Prinzipien geregelt ist, erleichtert werden. Diese Prinzipien werden im Folgenden dargestellt.

Eine zentrale Stellung kommt dem silbischen Prinzip zu, das im amtlichen Regelwerk keine eigenständige Rolle spielt. Nach Hinney (1997) stellt es den „missing link" zwischen dem phonographischen und dem morphologischen Prinzip her. Der Aufbau der Schreibsilbe wird in Abbildung 5.1 veranschaulicht.

Abbildung 5.1: Struktur der Schreibsilbe



(Blatt & Jarsinski, 2009, S. 93)

Jede Schreibsilbe hat einen verbindlichen Silbenkern und einen möglichen Silbenanfangs- und -endrand. Prototypisch für die deutsche Sprache ist der Zweisilber mit einer Abfolge von betonter und unbetonter Silbe (Trochäus), z. B. T**a**-fel und T**a**n-te. Die Typenbildung der Wortschreibungen beruht auf der regelhaften Struktur des Zweisilbers im Kernbereich und ist in Tabelle 5.1 dargestellt (vgl. Blatt, 2010, S. 104 f.).

Tabelle 5.1: Vier Grundtypen des prototypischen Zweisilbers im Kernbereich

	Offen	Geschlossen
Unmarkiert	Tafel	Tante
Markiert	Liebe	Tanne

(Blatt & Jarsinski, 2009, S. 93)

Die daran orientierte Aussprache der Vokale als Silbenkerne ist in der betonten Silbe entweder lang (T**a**-fel) oder kurz (T**a**n-te). Dies hängt davon ab, ob es sich um eine offene oder geschlossene Schreibsilbe handelt, je nachdem ob ein Silbenendrand vorhanden ist oder nicht. Da die betonte Silbe in T**a**-fel keinen Endrand aufweist, sondern auf dem vokalischen Silbenkern endet, wird das /a/ lang gesprochen. Die Silbe T**a**n-te verfügt über einen Endrand <n> und endet somit auf einem Konsonanten, weshalb das /a/ kurz gesprochen wird.

Die Vokallänge lässt sich also aus dem Silbenschnitt ableiten. Sie wird mit Ausnahme des Graphems <ie> nicht regelhaft markiert. Bei der geschlossenen Silbe wird die Vokalkürze in den Fällen regelhaft durch die Verdopplung des Konsonantenbuchstabens (Ta**nn**e) markiert, in denen eine sogenannte Silbengelenkschreibung vorliegt, d. h., wenn im Silbenschnitt nur ein Konsonant vorkommt (Eisenberg, 1995). Hier zeigt sich der Einfluss des silbischen Prinzips auf die Aussprache, die anders als im amtlichen Regelwerk angenommen nicht nur durch die Phonem-Graphem-Korrespondenz geregelt wird.

Zum morphologischen Prinzip zählen die Umlautschreibung, z. B. <hält>, weil <halten>, die entweder vom Ein- zum Zweisilber oder umgekehrt erschlossen werden kann, die Auslautverhärtung (z. B. Hund-Hunde) und die vererbten Silbengelenkschreibungen, z. B. <stellt>, weil <stellen>. Weiterhin zählen Flexionsmorpheme wie Personalendungen von Verben dazu.

Ableitung und Komposition sind in der deutschen Sprache weit verbreitet. Sie zählen zum Prinzip der Wortbildung, das ebenfalls zur Morphologie gehört. Ableitungen werden mithilfe von Prä- und Suffixen gebildet, wodurch eine neue Wortbedeutung entsteht oder die Wortart verändert wird (Eisenberg, 2004, S. 247 ff.). Die Komposition hat eine zentrale Funktion bei der Wortschatzbildung im Deutschen (ebd., S. 226 f.).

Neben diesen den Kernbereich der Rechtschreibung betreffenden Prinzipien gibt es einen Peripheriebereich der Rechtschreibung, zu dem nicht regelhaft herleitbare Schreibungen wie Wörter mit Dehnungs-h oder Fremdwörter gehören (vgl. Blatt, 2010, S. 109).

Darüber hinaus weist die Graphematik das wortübergreifende Prinzip aus, das die Großschreibung, die dass-Schreibung und die Zeichensetzung regelt.

Die dargestellte Schriftsystematik wurde erstmals im Rahmen der IGLU-Ergänzungsstudie 2006 genutzt, um ein theoretisches Rahmenkonzept als Grundlage für einen sprachsystematischen Rechtschreibtest zu entwickeln und darauf aufbauend ein Rechtschreibkompetenzmodell differenziell zu modellieren. Es wird im folgenden Kapitel im Zusammenhang mit dem SRT dargestellt.

5.2 SPRACHSYSTEMATISCHER RECHTSCHREIBTEST

In Tabelle 5.2 wird das theoretische Rahmenkonzept aufgezeigt, das als Basis für die Testentwicklung des sprachsystematischen Rechtschreibtests (SRT) dient.

Tabelle 5.2: Rahmenkonzeption zum sprachsystematischen Rechtschreibtest

Orientierung an Prinzipien	Teilkompetenzen
Phonographisch-silbischer Kernbereich	Bezug herstellen zwischen Schrift- und Lautstruktur unter Berücksichtigung der silbenstrukturellen Informationen (Silbenanfangs- und -endrand und Silbenschnitt)
Morphologischer Kernbereich	Vererbte silbenschriftliche Informationen in flektierten und abgeleiteten Formen herleiten; Flexionsmorpheme kennen und anwenden
Peripheriebereich	Markierungen in offenen Silben setzen und vererbte Schreibweisen herleiten; Transfer bei Sonderfällen und Lernwörtern; Fremdwortschreibung
Prinzipien der Wortbildung	Wortarten und Wortbildungsmorpheme kennen und in Ableitungen und Komposita produktiv anwenden
Wortübergreifendes Prinzip	Syntaxstrukturen kennen und für Groß-, Getrennt- und Zusammenschreibung, dass-Schreibung und Kommasetzung anwenden

(vgl. Blatt, 2010, S. 108 f.)

Das Rahmenkonzept weist fünf Teilkompetenzen auf, deren Inhalte sich an den graphematischen Forschungsergebnissen ausrichten (Blatt et al., 2011; Voss et al., 2007). Der SRT liegt in unterschiedlichen Testformaten vor, und zwar als Ganztext, als Wörtertest, als Satztest sowie als Lückentest und als Kombination von Testformaten. Bei der Entwicklung des Tests für unterschiedliche Klassenstufen werden Wörter ausgewählt, die die Lerninhalte der jeweiligen Rahmenpläne repräsentieren. Die Wörter werden in Struktureinheiten zerlegt, die den einzelnen Teilkompetenzen zugeordnet werden. Dabei wird darauf geachtet, dass in jeder Teilkompetenz eine für die Skalierung ausreichende Anzahl von Items enthalten ist. Die Auswertung wird auf der Ebene des ganzen Wortes und auf der Grundlage der Struktureinheiten vorgenommen. Die Auswertungsergebnisse auf Strukturebene liefern differenzierte Aussagen über die Rechtschreibkompetenz der Schülerinnen und Schüler und können somit als verlässliche

Grundlage zur Erstellung eines individuellen Förderplanes dienen. Exemplarisch ist in Tabelle 5.3 die Zuordnung der Struktureinheiten auf die fünf Prinzipien veranschaulicht.

Tabelle 5.3: Beispiel der Struktureinheiten zu den fünf Prinzipien

Orientierung an den Prinzipien	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
Zuordnungen	Hals**kette** #kette	**Hals**kette #Hals		Halskette #Kompositum	**H**alskette #H
	Wolkenmeer #Wolke	Wolk**en**meer #en	Wolken**meer** #meer	Wolkenmeer #Kompositum	**W**olkenmeer #W
		Zirkus**zelt** #zelt	**Zirkus**zelt #Zirkus	Zirkuszelt #Kompositum	**Z**irkuszelt #Z
	Ver**gangen**heit #gangen			**Ver**gangen**heit** #Ver #heit	**V**ergangenheit #V
		ents**chuld**igen #en #schuld		**ent**schuld**ig**en #ent #ig	
	wunderbar #wunder				

Der SRT wurde in den Ergänzungsstudien Orthografie zu IGLU 2006 (IGLU-E) und zu HeLp 2007/08 (HeLp-E) sowie in mehreren Studien des Nationalen Bildungspanels (NEPS) eingesetzt. Auf der Datengrundlage der Entwicklungsstudie in Klassenstufe 5 wurde der Test auf seine Testeigenschaften hin untersucht und es wurden unterschiedlich komplexe Modellierungen der Rechtschreibkompetenz vorgenommen mit dem Ziel, dass Kompetenzmodell empirisch zu untersuchen. Im Folgenden werden die Ergebnisse aus der IGLU-E-Vorstudie und der HeLp-E-Studie zusammengefasst (vgl. Blatt et al., 2011).

TESTFORMATE

Bei der Anwendung des SRTs in den unterschiedlichen Studien wurden verschiedene Testformate erprobt. In der IGLU-E-Studie 2006 wurde ein zusammenhängender Text eingesetzt,

wodurch ein syntaktischer und textueller Kontext der Testwörter gegeben war. Allerdings war ein hoher Zeitaufwand damit verbunden und aufgrund des geschlossenen Textes wiesen viele Testwörter eine hohe Lösungswahrscheinlichkeit auf und mussten von den Analysen ausgeschlossen werden. Aus diesem Grund wurde in der längsschnittlichen HeLp-E-Studie das Testformat verändert, um Tests für den Längsschnitt zu erproben. Zunächst wurde ein Wortdiktat mit 20 bzw. 30 Einzelwörtern zum ersten und zweiten Messzeitpunkt eingesetzt. Für den dritten Messzeitpunkt wurde eine Kombination aus Einzelwörtern und drei diktierten Sätzen gewählt, da mit dem Wortdiktat weder die Großschreibung noch die Getrennt- und Zusammenschreibung adäquat gemessen werden konnte. Diese Kombination ermöglichte eine zeitökonomische und zuverlässige Testung der Rechtschreibleistung.

ITEM- UND PERSONENPARAMETER

Mithilfe von IRT-Verfahren wurden die Item- und Personenparameter aus den empirischen Daten geschätzt. Auf dieser Grundlage lässt sich die Eignung der Testwörter (Item) in Form des Item-Fits und die Passung der Itemschwierigkeit mit der Personenfähigkeit überprüfen.

Die Ergebnisse der Datenanalyse für die IGLU-E-Vorstudie haben ergeben, dass die Item-Fits im Bereich zwischen 0.8 und 1.2 liegen und somit die Rechtschreibkompetenz mit den Items kohärent gemessen werden kann. Die Itemschwierigkeit der Testwörter steht in einem angemessenen Verhältnis zur gemessen Personenfähigkeit, d. h., dass der Test sowohl für gute als auch weniger gute Schülerinnen und Schüler Items beinhaltet.

Bezüglich der Item-Fits hat die längsschnittliche Datenanalyse der HeLp-E-Studie ergeben, dass sich die meisten Items im Bereich von 0.8 und 1.2 verteilen und nur wenige Items von der Analyse ausgeschlossen werden mussten. Demnach ist der Test in der Lage, die Rechtschreibkompetenz in kohärenter Weise zu messen, was sich zudem in der Passung der Itemschwierigkeit und Personenfähigkeit bestätigte.

5.2.1.1 EMPIRISCHE VALIDIERUNG DES KOMPETENZMODELLS IN DER IGLU-E-STUDIE

Die Überprüfung des SRT-Kompetenzmodells fand im Rahmen der Internationalen Grundschul-Leseuntersuchung (IGLU) als Ergänzungsstudie mit insgesamt drei Untersuchungen für die 4. Jahrgangsstufe statt. Als Testformat wurde ein Diktat gewählt und für die Analyse werden Struktureinheiten aus den Testwörtern abgeleitet.

Tabelle 5.4: Stichproben der IGLU-E-Studie zum SRT

Studie	N	Struktureinheiten
Voruntersuchung	486	170
Vergleichsuntersuchung	562	256
Hauptuntersuchung	2557	?[6]

In der stattgefundenen Voruntersuchung zu IGLU-E (N = 486) wurde der Test pilotiert. Er enthielt 170 Struktureinheiten für die Überprüfung des Kompetenzmodells. In der sogenannten IGLU-Vergleichsuntersuchung (N = 562) wurden drei Rechtschreibtests (SRT, DSP, DoSe) eingesetzt. Der dort eingesetzte SRT umfasste 256 Struktureinheiten. Die Hauptuntersuchung entspricht dem Design der Vergleichsstudie und basierte lediglich auf einer größeren Stichprobe (N = 2.557). Die Ergebnisse der Hauptuntersuchung zum SRT liegen noch nicht vor, weshalb für die Überprüfung des Kompetenzmodells auf die Vor- und Vergleichsuntersuchung zurückgegriffen wird.

In Abbildung 5.2 sind die Reliabilitäten und latenten Korrelationen (intra-domain-correlations) der Teilkompetenzen in der unteren Dreiecksmatrix für die Voruntersuchung und in der oberen Dreiecksmatrix für die Vergleichsuntersuchung dargestellt.

[6] Die Ergebnisse aus der Hauptuntersuchung der IGLU-E-Studie 2006 sind noch nicht veröffentlicht worden.

Abbildung 5.2: Latente Korrelationen und Reliabilitäten der fünf Teilkompetenzen in der IGLU-Vor- und Vergleichsuntersuchung

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
Phonographisch-silbischer Kernbereich	0.88 0.90	0.97	0.84	0.96	0.83
Morphologischer Kernbereich	0.95	0.90 0.91	0.86	0.95	0.81
Peripheriebereich	0.73	0.81	0.87 0.86	0.79	0.70
Prinzip der Wortbildung	0.91	0.91	0.82	0.86 0.89	0.88
Wortübergreifendes Prinzip	0.80	0.81	0.69	0.89	0.84 0.86

Latente Interkorrelationen (dargestellt auf der oberen und unteren Dreiecksmatrix) sowie Reliabilitäten (dargestellt auf der Diagonalen) der fünf Teilfähigkeiten in der Vergleichsuntersuchung (obere Angaben) und der Voruntersuchung (untere Angabe)

(Blatt et al., 2011, S. 246)

In beiden Untersuchungen konnten auf der Ebene der Teilkompetenzen Reliabilitäten über 0.80 festgestellt werden, was die zuverlässige Erfassung der Rechtschreibleistung im Sinne des differenziellen Rechtschreibkompetenzmodells mit dem SRT gewährleistet. Mit Ausnahme des Peripheriebereichs sind die Zusammenhänge der Teilkompetenzen für die Vor- und Vergleichsuntersuchung ähnlich. Ein sehr hoher Zusammenhang (0.95/0.97) besteht demnach bei den beiden Teilkompetenzen des Kernbereichs, also zwischen dem phonographisch-silbischen und morphologischen Prinzip. In Bezug auf den Kernbereich weist der Peripheriebereich Unterschiede bei der Vor- und Vergleichsuntersuchung auf. So liegen die Korrelationen zwischen dem phonographisch-silbischen Prinzip und dem Peripheriebereich bei der Voruntersuchung bei 0.73 und bei der Vergleichsuntersuchung bei 0.84. Moderater fällt der Unterschied zwischen dem morphologischen Prinzip und dem Peripheriebereich mit 0.81 bzw. 0.86 bei der Vor- und Vergleichsuntersuchung aus. Ansonsten sind die Korrelationen der Teilkompetenzen bei den beiden Studien ähnlich.

Die Modellpassung hat ergeben, dass die Repräsentation der Datenstruktur durch das mehrdimensionale Modell mit einem signifikanten Unterschied der Deviance-Werte von 1065 Punkten bei der Voruntersuchung und von 777 Punkten bei der Vergleichsuntersuchung besser als bei einem eindimensionalen Modell gegeben ist.

Zusammenfassend konnte in beiden Untersuchungen im Rahmen der IGLU-E-Studie die empirische Validierung des Kompetenzmodells für die 4. Jahrgangsstufe erzielt werden. Folglich ist der SRT geeignet, die Rechtschreibkompetenz der Schülerinnen und Schüler im Sinne einer kompetenzorientierten Leistungsmessung abzubilden. Trotz der hohen Korrelationen zwischen den Teilkompetenzen ist davon auszugehen, „dass das Konstrukt der Rechtschreibkompetenz bei Kindern der vierten Klassenstufe in sich differenziert ist“ (Blatt et al., 2011, S. 247).

5.2.1.2 EMPIRISCHE VALIDIERUNG DES KOMPETENZMODELLS IN DER HELP-STUDIE

Die Überprüfung des Kompetenzmodells des SRTs fand im Rahmen des Hamburger Leseförderprojekts (HeLp) längsschnittlich zu drei Messzeitpunkten für die 5. Jahrgangsstufe statt. Aus zeitökonomischen Gründen wurde das Testformat verändert. Die in den Tests enthaltenen Struktureinheiten für die Analyse sind in der Tabelle 5.5 für die drei Messzeitpunkte veranschaulicht.

Tabelle 5.5: Stichproben der HeLp-E-Studie zum SRT

Messzeitpunkt	N	Struktureinheiten
1 – September 2007	603	61
2 – Januar 2008	579	89
3 – Juli 2008	580	169

Zum ersten Messzeitpunkt (N = 603) umfasste der Test 20 Einzelwörter, aus denen 61 Struktureinheiten für die Überprüfung des Kompetenzmodells abgeleitet wurden. Weitere 10 Testwörter enthielt der Test zum zweiten Messzeitpunkt (N = 579), wodurch insgesamt 89 Struk-

tureinheiten für die Analyse bereitstanden. Der letzte Messzeitpunkt (N = 580) umfasste neben dem Wortdiktat auch drei Sätze mit insgesamt 28 Wörtern, aus denen sich 169 Struktureinheiten ergaben.

In Abbildung 5.3 sind die Reliabilitäten und latenten Korrelationen (intra-domain-correlations) der Teilkompetenzen in der oberen Dreiecksmatrix für den ersten Messzeitpunkt und in der unteren grau hervorgehobenen Dreiecksmatrix für den dritten Messzeitpunkt dargestellt.

Abbildung 5.3: Latente Korrelationen und Reliabilitäten der fünf Teilkompetenzen in der ersten und dritten HeLp-Erhebung

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
Phonographisch-silbischer Kernbereich	0.90 0.94	0.96	0.92	0.93	0.34
Morphologischer Kernbereich	0.99	0.91 0.94	0.94	0.90	0.43
Peripheriebereich	0.97	0.98	0.87 0.93	0.84	0.26
Prinzip der Wortbildung	0.98	0.98	0.93	0.88 0.91	0.30
Wortübergreifendes Prinzip	0.76	0.75	0.73	0.82	0,74 0.74

Latente Interkorrelationen (dargestellt auf der oberen und unteren Dreiecksmatrix) sowie Reliabilitäten (dargestellt auf der Diagonalen) der fünf Teilfähigkeiten zum ersten Testzeitpunkt der HeLp-Studie (obere Angaben) und dem dritten Testzeitpunkt (untere Angabe)

(Blatt et al., 2011, S. 248)

Mit Ausnahme des wortübergreifenden Prinzips konnten für beide Messzeitpunkte Reliabilitäten über 0.80 für die Teilkompetenzen nachgewiesen werden. Mit einer Reliabilität von jeweils 0.74 beim wortübergreifenden Prinzip zum ersten und dritten Messzeitpunkt kann noch davon ausgegangen werden, dass eine akzeptable Erfassung der Teilkompetenz erfolgt. Die niedrigen Werte sind vermutlich dem Testformat geschuldet, da bei den Einzelwörtern die Groß- und

Kleinschreibung nicht valide erfasst werden kann. Trotzdem ist eine zuverlässige Erfassung der Rechtschreibkompetenz gewährleistet.

Die Korrelationen sind wie bei der IGLU-E-Studie auf einem hohen Niveau, mit Ausnahme des wortübergreifenden Prinzips. Zum ersten Messzeitpunkt liegen die Zusammenhänge des wortübergreifenden Prinzips zu den anderen Teilkompetenzen zwischen 0.26 und 0.43. Die Werte sind „erwartungswidrig niedrig und drücken einen fehlenden Zusammenhang zu den übrigen Teilkompetenzen aus“ (Blatt et al., 2011, S. 248). Auch dieses Ergebnis kann auf das Testformat zurückgeführt werden. Auf einem vergleichbaren Niveau zur IGLU-E-Studie befinden sich die Korrelationen des wortübergreifenden Prinzips zum dritten Messzeitpunkt mit einem Korrelationskoeffizienten zwischen 0.73 und 0.82.

Die Modellpassung hat ergeben, dass die Repräsentation der Datenstruktur durch das mehrdimensionale Modell mit einem signifikanten Unterschied der Deviance-Werte von 604 Punkten zum ersten Messzeitpunkt und von 432 Punkten zum dritten Messzeitpunkt besser als bei einem eindimensionalen Modell gegeben ist.

Zusammenfassend konnte auch in der 5. Jahrgangsstufe die empirische Validierung des Kompetenzmodells gezeigt werden. Der SRT ist somit geeignet, die Rechtschreibkompetenz der Schülerinnen und Schüler in der 4. und 5. Jahrgangsstufe abzubilden (vgl. ebd., S. 247).

Für die weiteren Jahrgangsstufen der Sekundarstufe erfolgt die empirische Überprüfung des Kompetenzmodells im Rahmen des Nationalen Bildungspanels (NEPS). Ein weiteres Testformat kam bei der NEPS-Studie zum Einsatz, das aus einer Kombination von einem Lückentext und ganzen Sätzen besteht, womit wieder ein Satzkontext hergestellt werden konnte. Gegenstand dieser Arbeit ist die Darstellung der Überprüfung von der 5. bis zur 7. Jahrgangstufe.

6. DOMÄNE RECHTSCHREIBUNG IN DER NEPS-STUDIE

Die in dieser Arbeit durchgeführte Untersuchung zur längsschnittlichen Erfassung der Rechtschreibkompetenz ist in der deutschen Längsschnittstudie Nationales Bildungspanel (National Educational Panel Study, NEPS) verankert. Inhalte und Struktur der NEPS-Studie werden zunächst dargestellt, um anschließend den Fokus auf die Erfassung und Modellierung der Rechtschreibkompetenz im Rahmen von NEPS zu lenken.

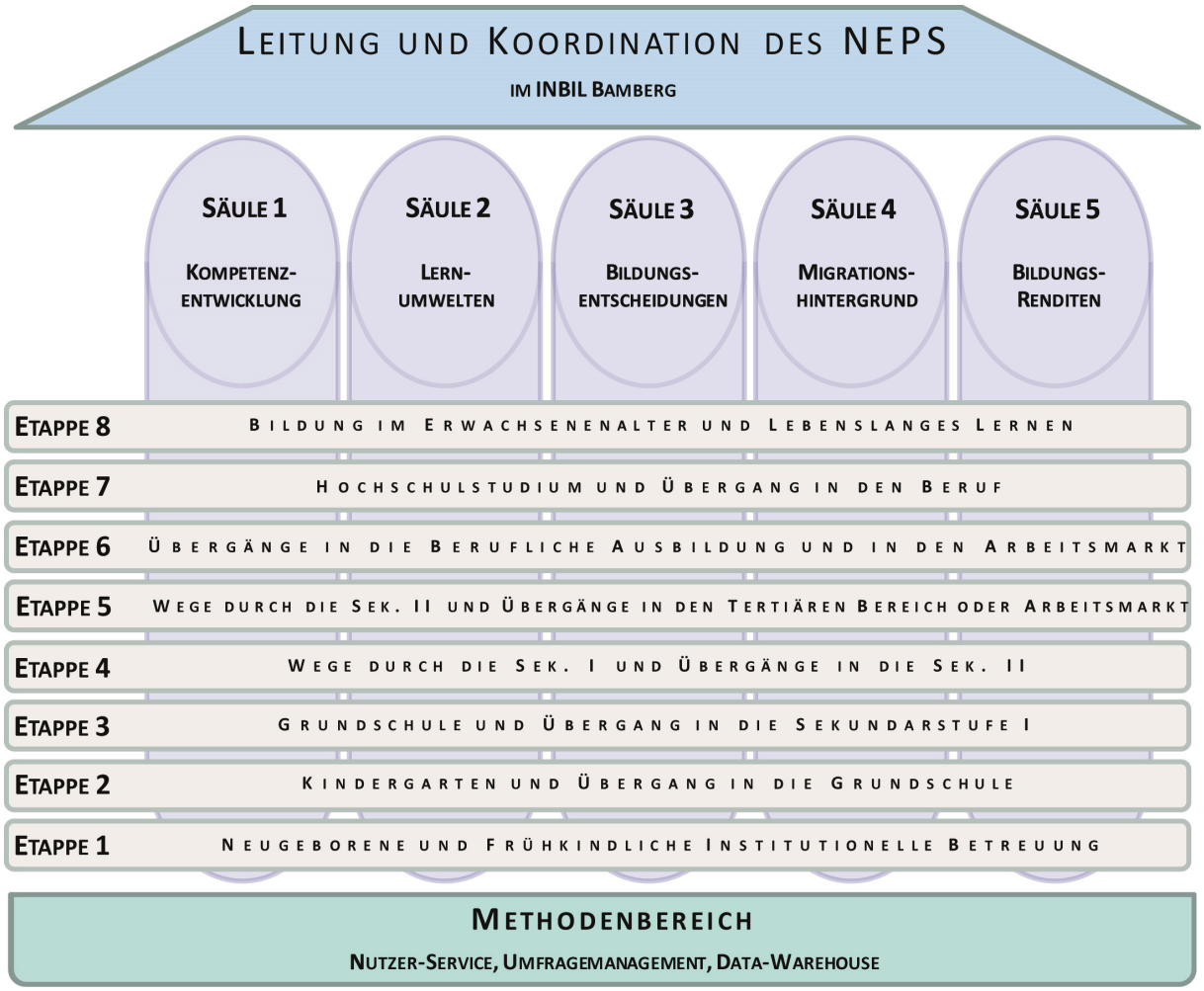
6.1 NEPS-STUDIE

NEPS war in der ersten Förderphase 2009 bis 2013 ein durch das Rahmenprogramm zur Förderung der empirischen Bildungsforschung des Bundesministeriums für Bildung und Forschung (BMBF) finanziertes Projekt. Das interdisziplinär zusammengesetzte Exzellenznetzwerk wurde bis Juli 2012 von Prof. Dr. Hans-Peter Blossfeld und wird seit August 2012 von Prof. Dr. Hans-Günther Roßbach geleitet. Die Leitung und Koordination sind am Institut für bildungswissenschaftliche Längsschnittforschung (INBIL) an der Otto-Friedrich-Universität Bamberg angesiedelt. Ab Januar 2014 wird die NEPS-Studie als Leibniz-Institut für Bildungsverläufe e.V. (LIfBi) fortgeführt.

NEPS verfolgt das Ziel, längsschnittlich „den Erwerb sowie die Konsequenzen von Bildung im Lebenslauf zu untersuchen und zentrale Bildungsprozesse und -verläufe über die gesamte Lebensspanne zu beschreiben“ (Blossfeld, von Maurice & Schneider, 2011b, S. 5). Dies bedeutet, dass die NEPS-Studie im Gegensatz zu den querschnittlichen Vergleichsstudien nicht nur Bildungsabschnitte betrachtet, sondern über die gesamte Lebensspanne Längsschnittdaten zur Kompetenzentwicklung, zu Bildungsprozessen, Bildungsentscheidungen und Bildungsrenditen in formalen, nicht-formalen und informellen Kontexten erhebt (vgl. Blossfeld, von Maurice & Schneider, 2011a, S. 2).

Die in Abbildung 6.1 dargestellte NEPS-Rahmenkonzeption umfasst die sogenannten Etappen und Säulen sowie einen Methodenbereich.

Abbildung 6.1: NEPS-Rahmenkonzeption



In acht Etappen werden die Bildungsbiografien untersucht, um längsschnittliche Informationen zu den Bildungsverläufen und -übergängen von den Neugeborenen (Etappe 1) bis zum Erwachsenenalter (Etappe 8) zu gewinnen. Die fünf Säulen sind verantwortlich dafür, in fünf theoretisch verbundenen Dimensionen Instrumente zu entwickeln bzw. auszuwählen und zur Verfügung zu stellen (vgl. Blossfeld et al., 2011a, S. 9 ff.):

- **Kompetenzentwicklung:** In Säule 1 stehen die Analyse von Kompetenzen und die Entwicklung von Testinstrumenten zur längsschnittlichen Messung der Kompetenzen im Fokus.
- **Lernumwelten:** Inwiefern Bildungsprozesse und der Kompetenzerwerb durch die jeweiligen Lernumwelten beeinflusst werden, liegt im Zuständigkeitsbereich von Säule 2.

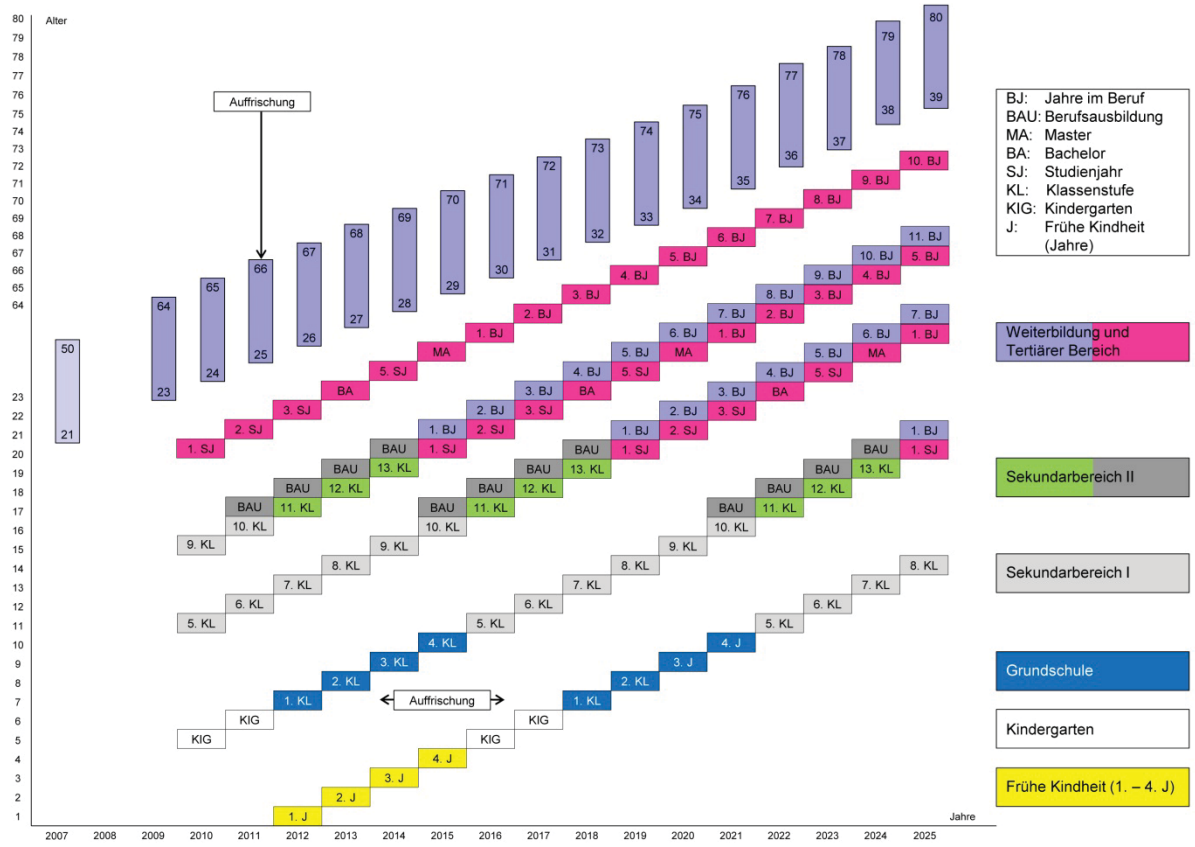
- **Bildungsentscheidungen:** Säule 3 beschäftigt sich mit sozioökonomischen- und geschlechtsspezifischen Unterschieden bei Bildungsentscheidungen, Bildungsaspirationen sowie der Fächerwahl.
- **Migrationshintergrund:** Welche Besonderheiten aufgrund einer Migrationsgeschichte beim Kompetenzerwerb und der Kompetenzentwicklung sowie bei den Bildungsentscheidungen auftreten, gehört in den Verantwortungsbereich von Säule 4.
- **Bildungsrenditen:** Die Erträge von Bildung in Form von ökonomischen, politischen, sozialen, physischen und psychischen Faktoren werden mit Instrumenten aus Säule 5 erfasst.

Neben den in den Etappen und Säulen definierten Zuständigkeiten bei der längsschnittlichen Erfassung der Bildungsverläufe besteht darüber hinaus die Möglichkeit, sogenannte etappenspezifische Ergänzungen durchzuführen, um besondere Bildungsprozesse in den Blick zu nehmen.

Zum NEPS-Methodenbereich gehören die Methodengruppe „Methoden, Gewichte und Imputation“, das Data Warehouse, das Technology Based Assessment (TBA) des Deutschen Instituts für internationale pädagogische Forschung in Frankfurt (DIPF), der Nutzer-Service sowie die Mode Effect Studies. Die Methodengruppe steht den Säulen und Etappen bei Fragen zur Stichprobenziehung und -gewichtung sowie der Imputation und statistischen Längsschnittauswertung beratend zur Seite. Das Data Warehouse ist für die Entwicklung von Datenbanken und Software für das NEPS verantwortlich, wohingegen sich das TBA um innovative Befragungs- und Testmethoden kümmert. Der Nutzer-Service stellt Informationen zum NEPS bereit und sorgt für eine benutzerfreundliche Datennutzung, indem z. B. Dokumentationen und Nutzerschulungen zu den Daten bereitgestellt werden. Im Rahmen der Mode Effect Studies unter der operativen Leitung von Dr. Ulf Kröhne werden im Hinblick auf technologiebasierte Testdurchführungen mögliche Unterschiede zwischen papierbasierten und computerbasierten Tests untersucht.

Der NEPS-Studie liegt ein Multi-Kohorten-Sequenz-Design zugrunde. Dies ermöglicht es, gleichzeitig unterschiedliche Gruppen bzw. Startkohorten zu bestimmten Lebens- und Bildungsphasen längsschnittlich zu begleiten. Fokussiert werden dabei die Übergänge im Bildungssystem, wozu der Eintritt in die Grundschule, die Übergänge in das gegliederte Schulsystem bzw. die berufliche Ausbildung oder das Studium und der Arbeitsmarkteintritt gehören (vgl. Abbildung 6.2).

Abbildung 6.2: Multi-Kohorten-Sequenz-Design des Nationalen Bildungspanels



Das NEPS-Design bezieht sich zunächst auf einen geplanten Testzeitraum von 2009 bis 2019, wobei sechs Startkohorten (Frühe Kindheit, Kindergarten, Grundschule, Sekundarbereich I & II, Weiterbildung und tertiärer Bereich) parallel längsschnittlich untersucht werden. Die Startkohorten decken eine Altersspanne von 1 bis 74 Jahren ab, wodurch Bildungsprozesse und -verläufe über die gesamte Lebensspanne im Rahmen von sogenannten Haupterhebungen (HE) erfasst und beschrieben werden. Im Vorfeld einer Haupterhebung werden die Instrumente und Abläufe in Entwicklungsstudien (ES) und Großpiloten (GP) erprobt. So erfolgte z. B. im Jahr 2010 die erste Haupterhebung mit den Startkohorten Kindergarten (KIG), Sekundarstufe I (5./9. KL) sowie den Studierenden (SJ) (vgl. Blossfeld et al., 2011a, S. 21 f.).

Für die sechs Startkohorten sind folgende Stichprobengrößen für die Teilnahme an den Befragungen und Testungen der NEPS-Studie vorgesehen.

Tabelle 6.1: Stichprobengröße des Nationalen Bildungspanels

Startkohorte	Stichprobengröße	Stichprobenbasis
Neugeborene	3.000	Individuum
Kindergarten ↳ *Grundschule*	3.000 *6.000*	Institution
Klassenstufe 5	6.800	Institution
Klassenstufe 9 ↳ *Gymn. Oberstufe* ↳ *Ausbildung / DS*	13.500	Institution
Studienanfänger (tertiärer Bereich)	16.500	Institution
Erwachsene	13.000	Individuum

(Aßmann et al., 2011)

Neben der Differenzierung der Startkohorten wird zwischen zwei Arten von Stichproben unterschieden, d. h., dass Daten entweder aus einer Individualstichprobe für einzelne Individuen (Neugeborene, Erwachsene) oder aus einer Klumpenstichprobe für eine Einheit bzw. Gruppe aus einer Institution (Kindergarten, Schule, Hochschule) erhoben werden. Die Teilnehmerzahlen liegen je nach Kohorte zwischen 3.000 und 16.500 Personen, wobei zum Teil (Frühe Kindheit, Kindergarten, Schule) das unmittelbare Umfeld mit in die Untersuchung eingebunden wird. Beispielsweise werden bei der Startkohorte in der Sekundarstufe (5./9. Klassenstufe) die Eltern der Schülerinnen und Schüler, die Lehrerinnen und Lehrer sowie die Schulleiterinnen und Schulleiter in die Untersuchung einbezogen. Im Hinblick auf eine zu erwartende Panelmortalität (vgl. Kapitel 1.2.1) bei der längsschnittlichen Erfassung der Bildungsverläufe werden individuelle Nachverfolgungen, Auffrischungen und Aufstockungen zur Panelpflege eingesetzt (vgl. Blossfeld et al., 2011b, S. 15 f.).

NEPS erhebt vielschichtige Längsschnittsdaten, die der Scientific Community für Forschungszwecke kostenlos zur Verfügung gestellt werden (vgl. Blossfeld et al. 2011a, S. 25 f.). Die Freigabe der Daten erfolgt 18 Monate nach Abschluss der Feldphase der jeweiligen Startkohorten und ist in einem bereitgestellten Zeitplan für alle Interessenten ersichtlich. Die NEPS-Daten werden durch das Datenzentrum fortlaufend erweitert und aktualisiert.

Wissenschaftlerinnen und Wissenschaftler können Anträge zur Nutzung von Daten stellen. Dabei werden das konkrete Forschungsvorhaben und -ziel skizziert sowie die Nutzungsdauer der Daten und die involvierten Mitarbeiterinnen und Mitarbeiter einer Institution festgelegt. Die Daten können auch für Qualifikationsarbeiten wie Dissertationen beantragt werden. Der Antrag ist beim NEPS-Datenzentrum einzureichen. Bei Bewilligung wird ein Datennutzungsvertrag geschlossen. Berechtigt ist jede Wissenschaftlerin und jeder Wissenschaftler einer Institution aus dem In- und Ausland.

Um der Scientific Community die Datennutzung ohne Vorkenntnisse zu ermöglichen, wurde an der Universität Bamberg eine Infrastruktur geschaffen, die ein ausgeklügeltes Unterstützungs- und Datensicherungssystem zur Verfügung stellt. Dazu zählen Nutzerschulungen in Bamberg sowie die Dokumentationen (Technical Reports) und Such- und Aufbereitungshilfen zu den einzelnen Studien und Daten. Ebenso wurde auf einen einfachen Datenzugang geachtet, der sich besonders durch moderne Technologien auszeichnet. Zu den Daten gibt es drei Zugangswege:

- Herausgabe von Scientific Use Files (SUF), die über die NEPS-Website zum Download bereitgestellt werden,
- Datennutzung über eine moderne Fernzugriffstechnologie (RemoteNEPS) sowie
- Datenzugriff im Rahmen von Gastaufenthalten vor Ort („On-Site“).

Diese Zugangswege unterscheiden sich im Grad der Anonymisierung der sensiblen Daten und gewährleisten alle die strenge Einhaltung der datenschutzrechtlichen Vorgaben:

- Der Download des Scientific Use Files weist die stärkste Form der Anonymisierung auf, indem individuell zuordenbare Informationen entfernt werden bzw. durch die Aggregation der Daten verloren gehen.
- Der Fernzugriff über den RemoteNEPS ermöglicht die verschlüsselte Onlinearbeit mit einem nur geringfügig anonymisierten Scientific Use Files für Analysezwecke. Die Nutzer erhalten geprüfte Outputs.
- Ein kontrollierter On-Site Zugriff besteht vor Ort in Bamberg und eignet sich besonders für die lokale Datenanalyse mit sensiblen Daten (z. B. Anzahl der Schülerinnen und Schüler, Angaben auf Landesebene, geografische Angaben, Angaben zum institutionellen Kontext).

Durch diese verschiedenen Möglichkeiten können unterschiedliche Analyseinteressen bedient werden.

6.2 RECHTSCHREIBTESTUNG ALS ETAPPENSPEZIFISCHE ERGÄNZUNG – RECHTSCHREIB-TEILSTUDIE

Die Rechtschreib-Teilstudie wird im Rahmen der NEPS-Studie durch die Etappe 4 als etappenspezifische Ergänzung durchgeführt und beinhaltet die kompetenzorientierte Erfassung der Rechtschreibleistung sowie die Modellierung der Rechtschreibkompetenz (vgl. auch im Folg.: Frahm et al., 2011). Zudem werden durch die Befragung der Schülerinnen und Schüler, Eltern sowie Lehrerinnen und Lehrer Informationen gewonnen, um Hintergrundvariablen und mögliche Einflussfaktoren zu ermitteln. Einleitend werden die Organisation und Schwerpunkte der etappenspezifischen Teilstudie erläutert, um nachfolgend die verantwortlichen Erhebungen zu skizzieren. Der eingesetzte NEPS-Rechtschreibtest wird in Kapitel 7.3 beschrieben und erläutert.

Etappe 4 „Wege durch die Sekundarstufe I und Übergänge in die Sekundarstufe II" der NEPS Studie, die von Prof. Dr. Wilfried Bos vom Institut für Schulentwicklungsforschung (IFS) der TU Dortmund in Kooperation mit Prof. Dr. Inge Blatt von der Universität Hamburg wissenschaftlich geleitet wird, bildet das Bindeglied zwischen den Etappen der Primarstufe und der Sekundarstufe II bzw. dem Eintritt in den Arbeitsmarkt.

Die etappenspezifischen Ergänzungen konzentrieren sich auf drei Bereiche, die auf dem literacy-Konzept beruhen und von besonderer Bedeutung für die Sekundarstufe I sind:

- **Reading Engagement**: Dieser Schwerpunkt resultiert aus der zentralen Rolle, die Lesen für den Lernerfolg in der Sekundarstufe I spielt, sowie den vorliegenden Ergebnissen, dass das Reading Engagement ein signifikanter Prädiktor für die Lesekompetenz darstellt, der soziale Benachteiligung ausgleichen kann. Zum Reading Engagement werden vom NEPS-Team in Dortmund Fragebogeninstrumente entwickelt.
- **Quality of instruction**: Im Hinblick auf die Bedeutung der Sekundarstufe I für die weitere Bildungskarriere sowohl im Hinblick auf einen erfolgreichen Schulabschluss als auch für den Übergang in die akademische bzw. berufliche Bildung ist die Unterrichtsqualität von zentraler Bedeutung. Etappe 4 setzt sich daher das Ziel, verlässliche Daten

für pädagogisch-didaktische und bildungspolitische Zwecke zu erheben. Zu diesem Schwerpunkt werden ebenfalls vom NEPS-Team in Dortmund Fragebogeninstrumente entwickelt.

– **Rechtschreibkompetenz**: Eine gute Rechtschreibkompetenz ist für den schulischen und beruflichen Erfolg wichtig, da sie zum Teil bereits eine Voraussetzung für einen Übergang in den Gymnasialzweig in Klassenstufe 5 bildet und die schulische Laufbahn bzw. die beruflichen Ausbildungschancen in hohem Maße bestimmt. Die längsschnittliche Entwicklung eines Rechtschreibtests sowie von Schüler-, Lehrer- und Elternfragebogeninstrumenten zur Rechtschreibung liegt in der Verantwortung der Kooperationspartner an der Universität Hamburg.

Mit dem Einsatz der entwickelten Instrumente im Verlauf der Sekundarstufe I leistet die Etappe 4 im Rahmen von NEPS erstmalig einen Beitrag zur nationalen längsschnittlichen Kompetenzerfassung über mehr als zwei Messzeitpunkte in den aufgezeigten sprachlichen Schwerpunktbereichen. Für diesen Zweck werden die Test- und Fragebogeninstrumente interdisziplinär aus fachdidaktischer und empirischer Sicht entwickelt und im Rahmen von Entwicklungsstudien (ES) und Großpiloten (GP) vor dem Einsatz in den Haupterhebungen (HE) erprobt. Erprobungen im Rahmen von Entwicklungsstudien bzw. Großpiloten werden jährlich durchgeführt. Tabelle 6.2 gibt einen Überblick über den bisherigen Einsatz von der 5. Klassenstufe (K5) bis zur 9. Klassenstufe (K9).

Tabelle 6.2: Übersicht der Entwicklungsstudien und Großpiloten der Etappe 4

Studie	Jahr	Klassenstufe	N
ES K5	Herbst 2009	5	298
ES K6	Frühjahr 2011	6	414 (307*)
ES K7	Herbst 2011	7	307 (307*)
GP K8	Herbst 2012	8	314
GP K9	Herbst 2013	9	[7]
* Längsschnitt der identischen Schülerinnen und Schüler			

[7] Aktuell in Erhebungsphase.

Für den Einsatz in den Haupterhebungen werden die Test- und Fragebogeninstrumente auf Basis der Ergebnisse aus den Datenanalysen der Entwicklungsstudien und Großpiloten im Hinblick auf die Gütekriterien optimiert bzw. die Testinstrumente längsschnittlich weiterentwickelt.

Die von der Etappe verantwortete Rechtschreibtestung und Befragung in den Haupterhebungen erfolgt im Abstand von zwei Jahren in den Klassenstufen 5 (K5), 7 (K7) und 9 (K9). Die Durchführung in K5 und K7 ist abgeschlossen (vgl. Tabelle 6.3). Die Erhebung für K9 ist für 2014 geplant.

Tabelle 6.3: Übersicht der Haupterhebungen der Etappe 4

Studie	Jahr	Klassenstufe	N
HE K5	Herbst 2011	5	4.989
HE K7	Herbst 2012	7	6.196[8]

Die von der Etappe verantworteten Befragungen richten sich an Schülerinnen und Schüler, Eltern und Deutschlehrer. Den Schülerinnen und Schülern werden Fragen zur Rechtschreibung, Unterrichtsqualität und zum Reading Engagement in Form von Paper and Pencil Interviews (PAPI) gestellt. Die unterrichtenden Deutschlehrerinnen und -lehrer werden ebenfalls mithilfe eines PAPI, die Eltern der Schülerinnen und Schüler dagegen per Computer Assisted Telephone Interviews (CATI) befragt.

Die Auswertung der Daten zielt zum einen auf die Erstellung eines Scientific Use Files (SUF). Für die HE K5 liegt der SUF inklusive eines Technical Reports bereits vor. Zum anderen werden Daten im Rahmen von Qualifikationsarbeiten, Publikationen und Vorträgen analysiert (z. B. Frahm, 2013; Frahm et al., 2011; Frahm & Jarsinski, 2010, 2011; Jarsinski & Frahm, 2012; Jarsinski, Frahm, Blatt, Bos & Kanders, in Druck; Jarsinski & Frahm 2012).

[8] Die Stichprobe der HE K5 wurde zur HE K7 nochmals im Sinne eines Längsschnitts untersucht und um 2158 Schülerinnen und Schüler aufgestockt.

II. EMPIRISCHE UNTERSUCHUNG

7. FORSCHUNGSVORHABEN

Das vorliegende Forschungsvorhaben hat in erster Linie eine methodische Ausrichtung. Es baut auf Forschungsergebnissen zur kompetenzorientierten Messung der Rechtschreibleistung mithilfe eines sprachsystematischen Rechtschreibtests (SRT) auf (Kapitel 5.2). Auf der Grundlage des theoretischen und empirisch validierten sprachsystematischen Rahmenkonzeptes wurden in interdisziplinärer Zusammenarbeit SRTs für die längsschnittliche Erfassung der Rechtschreibkompetenz in der NEPS-Studie im Verlauf der Sekundarstufe I entwickelt und in Entwicklungsstudien erprobt (Kapitel 6.2). Dabei tauchen grundsätzliche Fragen auf, denen sich die vorliegende Arbeit widmet:

- Haben die in einer Entwicklungsstudie nachgewiesenen Gütekriterien für das Messinstrument SRT auch in der folgenden repräsentativen Haupterhebung Bestand?
- Welche Analyseverfahren eignen sich für die längsschnittliche Kompetenzentwicklung, um diese verlässlich und effizient zu erfassen und mögliche Änderungen der differenziellen Kompetenzstruktur aufzudecken?
- Welche Faktoren wirken sich auf die differenzielle Kompetenzentwicklung aus?

Als Datenbasis werden die Kompetenz- sowie ausgewählte Befragungsdaten aus den NEPS-Entwicklungsstudien in Klassenstufe 5 bis 7 (ES K5-K7) sowie der Haupterhebung in Klassenstufe 5 (HE K5) herangezogen.

7.1 FORSCHUNGSFRAGEN UND METHODISCHES VORGEHEN

Im Hinblick auf den Forschungsstand zur längsschnittlichen kompetenzorientierten Messung der Rechtschreibleistung (Kapitel 2.3) werden für die Untersuchung keine Hypothesen, sondern Forschungsfragen formuliert. Die Fragen werden zunächst dargestellt und kurz erläutert. Anschließend wird das methodische Vorgehen zur Beantwortung der Forschungsfragen vorgestellt. Abschließend wird eine Übersicht geliefert, in der die Forschungsfragen, die jeweilige Datengrundlage sowie die Auswertungsverfahren zusammengestellt sind.

FORSCHUNGSFRAGEN

Die Forschungsfragen beziehen sich auf die Testentwicklung, die Erfassung der Kompetenzentwicklung und -struktur sowie die Ermittlung von Einflussfaktoren:

F1: Wie verhalten sich die in der Entwicklungsstudie und in der Haupterhebung in Klassenstufe 5 ermittelten Gütekriterien für den Rechtschreibtest zueinander?

Die Stichproben von Entwicklungsstudie und Haupterhebung unterscheiden sich hinsichtlich der Größe und Repräsentativität. Hier soll herausgefunden werden, ob die Erprobung von neuen Testinstrumenten in kleinen Studien zu verlässlichen Ergebnissen im Hinblick auf die Testgüte führt.

F2: Mit welchen methodischen Verfahren lässt sich die Entwicklung der Rechtschreibkompetenz verlässlich und effizient erfassen?

Hier geht es darum, Verfahren zur längsschnittlichen Messung der Rechtschreibkompetenz mit den Daten der Entwicklungsstudien in Klassenstufe 6 und 7 vergleichend zu testen, um ihre Eignung zu untersuchen.

F3: Verändert sich die Kompetenzstruktur von Klassenstufe 6 bis 7, und wenn ja, in welcher Weise?

Neben der kompetenzorientierten Erfassung der Leistungsentwicklung stellt sich darüber hinaus die Frage, ob sich auch die Struktur der Rechtschreibkompetenz mit zunehmender Lernentwicklung ändert, und wenn ja, wie. Konkret wird mit den Daten der Entwicklungsstudien in Klassenstufe 6 und 7 untersucht, ob die dem Kompetenzmodell zugrundeliegende differenzielle Struktur von fünf Teilkompetenzen erhalten bleibt bzw. sich verändert.

F4: Beeinflussen die Entwicklung in den Teilkompetenzen bzw. die Faktoren Geschlecht, Alter, sozioökonomischer Hintergrund, Migrationshintergrund und Schulform die Entwicklung der Rechtschreibkompetenz auf der Ebene des ganzen Wortes von Klassenstufe 6 und 7, und wenn ja, in welcher Weise?

Hier wird untersucht, ob die Teilkompetenzen des differenziellen Rechtschreibkompetenzmodells bzw. die in den Schulleistungsstudien als relevant erkannten Einflussfaktoren auch die Entwicklung der Rechtschreibkompetenz von Klassenstufe 6 bis 7 beeinflussen, und wenn ja, wie.

METHODISCHES VORGEHEN

F1: Wie verhalten sich die in der Entwicklungsstudie und in der Haupterhebung in Klassenstufe 5 ermittelten Gütekriterien für den Rechtschreibtest zueinander?

Um diese Frage zu beantworten, erfolgte zunächst die Skalierung der Leistungsdaten der Schülerinnen und Schüler der Entwicklungsstudie (ES) und der Haupterhebung (HE) in Klassenstufe 5 (K5) mit der Skalierungssoftware ConQuest 2.0. Auf diese Weise wurden die zentralen Item-, Personen- und Modellparameter bestimmt. Dazu zählen bei den Itemparametern die Itemschwierigkeit, der Item-Fit und die Trennschärfe. Bei den Personenparametern ist die geschätzte Personenfähigkeit ausschlaggebend, wofür der WLE-Schätzer genutzt wurde, da er sich laut der Forschungsergebnisse zur Veränderungsmessung und aus Längsschnittstudien im Gegensatz zu den EAP- und PV-Schätzern am besten für diesen Zweck eignet (vgl. Hartig & Kühnbach, 2006, S. 42). Die Personenfähigkeit wurde anschließend auf einen Mittelwert von 500 und eine Standardabweichung von 100 normiert, indem die vorliegende Personenfähigkeit in Logits nach der Standardisierung bzw. z-Transformation mit dem Wert 100 multipliziert und der Wert 500 dazu addiert wurden (vgl. Eggert et al., 2010, S. 307 ff.). Die relevanten Modellparameter umfassen die Reliabilität, die Korrelation sowie die Modellgeltungstests zur Prüfung der Dimensionalität des Kompetenzkonstrukts. Da diese Parameter einem Messfehler unterliegen, der zur Beeinflussung der Ergebnisse bzw. Reduzierung der Aussagekraft der statistischen Tests führen kann, wird die Kompetenz latent auf der Grundlage aller zur Verfügung stehender Lücken und Sätze modelliert. Dies trägt zur Relativierung des Messfehlers bei. Zudem wird durch die Verwendung des WLE-Schätzers eine Korrektur der Varianz der Personenfähigkeit vorgenommen, die die Verzerrung des Personenparameters reduziert. Dennoch bleibt ein unabdingbarer Anteil des Messfehlers bestehen. Anhand von vier Bereichen werden die Ergebnisse auf Grundlage der Item-, Personen- und Modellparameter berichtet:

1. Datenprüfung
2. Eindimensionale Skalierungen
3. Mehrdimensionale Skalierungen
4. Dimensionalität des Kompetenzkonstrukts

Konkret wurden für diese Berechnungen in einem ersten Schritt die Daten der ES K5 herangezogen und einleitend wurde eine Datenprüfung für das ganze Wort mit der Statistiksoftware

SPSS 22 vorgenommen, indem die Normalverteilung[9] der Personenparameter sowie die Annahme des Rasch-Modells überprüft wurden. Dazu wurden die Annahme einer Normalverteilung statistisch mit dem Kolmogorov-Smirnov-Test getestet und zusätzlich wurden die geschätzten Personenfähigkeiten grafisch mit der Normalverteilungskurve in einem Histogramm dargestellt, um eine Verzerrung der Personenparameter auszuschließen. Zudem wurde die Messinvarianz zur Annahme des Rasch-Modells anhand eines grafischen Modelltests im Hinblick darauf überprüft, ob sich Abweichungen zwischen zwei Teilstichproben in Bezug auf die homogene Messung der Kompetenz ergeben. Die Bildung der Teilstichproben erfolgte unter Zuhilfenahme der jeweiligen Personenfähigkeiten der Schülerinnen und Schüler, indem zwei Gruppen mit einer Personenfähigkeit unter- und oberhalb eines WLE-Werts von 500 gebildet wurden, um so die jeweiligen Itemschwierigkeiten hinsichtlich der zwei Teilstichproben gegenüberzustellen.

Auf ein- und mehrdimensionaler Ebene schloss sich eine umfassende Analyse der Item-, Personen und Modellparameter für das ganze Wort und der fünf Teilkompetenzen an. Ferner wurde zwischen sogenannten Ursprungs- und Ausgangsmodellen unterschieden, wobei die deskriptive Analyse dieser Ursprungsmodelle mit allen Analyseeinheiten in Form der ganzen Wörter bzw. Struktureinheiten durchgeführt wurde, um die Ergebnisse zur Optimierung und Findung der Ausgangsmodelle zu nutzen.[10] Im Zuge dessen wurden auffällige Wörter und Struktureinheiten identifiziert, die nicht der Referenz für den Item-Fit (≤ 1.20) und der Trennschärfe (> 0.25) entsprachen und daher von den Analysen ausgeschlossen wurden. Zu den ermittelten Statistiken zählen bei der eindimensionalen Skalierung die absolute und relative Häufigkeit der leichten und schweren Wörter bzw. Struktureinheiten, der Mittelwert sowie der Standardfehler der Itemschwierigkeit[11] und Personenfähigkeit, die Reliabilität sowie die prozentuale Lösungshäufigkeit der Wörter und Struktureinheiten. Die mehrdimensionale Skalie-

[9] Die Überprüfung der Normalverteilung wurde aufgrund eines Softwarefehlers in der Statistiksoftware SPSS 22 mit dem kostenlosen Äquivalent PSPP vorgenommen, das einen vergleichbaren Analyseumfang bietet und zu identischen Ergebnissen führt (www.gnu.org/software/pspp/).

[10] Bis das optimale Skalierungsmodell gefunden wurde, waren mehrere Skalierungsläufe notwendig, da immer wieder Wörter bzw. Einheiten auffällig wurden. Diese Veränderungen können allerdings nicht alle aufgezeigt werden, weshalb sich in dieser Arbeit nur auf das Ursprungs- und Ausgangsmodell bezogen wird. Diese Vergleichsform bringt ein generelles durch die Software bedingtes Problem mit sich, und zwar, dass sich die Nummerierungen der jeweiligen Wörter und Struktureinheiten durch die Optimierung der Modelle ändern und somit ein 1:1-Abgleich nur bedingt möglich ist.

[11] Wegen der gleichbleibenden Zentrierung der Itemschwierigkeit bei der Skalierung durch die Restriktion „contraints = items“ resultiert ein Mittelwert von 0,00.

rung umfasst zudem noch die latenten Korrelationen und die Modellpassung zur Überprüfung der Dimensionalität des Kompetenzkonstrukts durch Modellgeltungstests. Dazu wurden die Deviance-Statistik, das Akaike's Information Criterion (AIC), der Consistent AIC (CAIC) sowie das Bayes Information Criterion (BIC) herangezogen. Die Modellierung des Kompetenzkonstrukts erfolgte in vier Varianten, um die fünf Teilkompetenzen empirisch zu modellieren. Bei dem fünfdimensionalen Modell nimmt jede Teilkompetenz eine Dimension ein. Die Zusammenführung des Kernbereichs ergibt das vierdimensionale Modell. Die weitere didaktisch und statistisch begründbare Reduzierung der Dimensionen führt zu einem zweidimensionalen Modell mit den wortbezogenen Prinzipien (1-4) und dem syntaxbezogenem Prinzip (5).

In einem zweiten Schritt wurden die Daten der HE K5 analog zu dem bereits beschriebenen Auswertungsvorhaben analysiert.

Im einem letzten Schritt wurde ein Vergleich mit der stichprobenmäßig größeren HE K5 vorgenommen, um zu untersuchen, ob die Erprobung des Testinstruments in der im Verhältnis kleineren ES K5 zu verlässlichen Ergebnissen führte. Stellvertretend für die weiteren Entwicklungsstudien und Haupterhebungen wurde mittels der vergleichbaren Ursprungsmodelle geprüft, inwiefern sich die Stichprobengröße und -zusammensetzung auf die Aussagekraft der Ergebnisse auswirkt.

Die Untersuchung und die Ergebnisse zu dieser Forschungsfrage werden in Kapitel 8.1 dargestellt.

F2: Mit welchen methodischen Verfahren lässt sich die Entwicklung der Rechtschreibkompetenz verlässlich und effizient erfassen?

Zur Beantwortung dieser Frage wurde in Anlehnung an Arbeiten aus anderen Domänen die Kompetenzmodellierung der Rechtschreibung auf der Ebene des ganzen Wortes anhand der drei in Kapitel 3.3.1 aufgezeigten Verfahren zur längsschnittlichen Kompetenzmodellierung für personenspezifische Veränderungen explorativ gegenübergestellt (vgl. Carstensen et al., 2012; Eggert et al., 2010). Damit sollte untersucht werden, wie sich die getrennte Skalierung (1) im Vergleich zu den Skalierungen mit virtuellen Personen (2) oder der Skalierung mit latenten Dimensionen (3) verhält. Zu diesem Zweck wurden die Möglichkeiten und Grenzen der drei Verfahren zur Kompetenzmodellierung der Rechtschreibung für den Längsschnitt der Entwicklungsstudien (ES) in Klassenstufe 6 (K6) und 7 (K7) mit 307 Schülerinnen und Schülern eruiert.

Das methodische Vorgehen zur Bestimmung der Item-, Personen- und Modellparameter orientiert sich an der Vorgehensweise, die zur Beantwortung der ersten Forschungsfrage dargestellt wurde. Für die Darstellung der drei Verfahren zur längsschnittlichen Kompetenzmodellierung wurde wegen der Übersichtlichkeit auf die Ergebnisse der ES K6 verzichtet, da die Personenfähigkeiten in allen Fällen auf einen Mittelwert von 500 normiert waren. Daher wird lediglich die Entwicklung der Rechtschreibkompetenz mit den Ergebnissen der ES K7 veranschaulicht, indem die WLE-Werte und Effektstärken nach Cohen berichtet und im Verhältnis zur prozentualen Lösungshäufigkeit eingeordnet werden. Zudem wurde bei der aufgezeigten längsschnittlichen Kompetenzermittlung zwischen den sogenannten Skalierungs- und Fixierungsvarianten unterschieden. Die Skalierungsvarianten zeigen auf, ob die Entwicklung der Kompetenz nur mit den 51 Ankeritems oder zusätzlich mit den sogenannten zeitpunktspezifischen Items ermittelt wurden (vgl. Tabelle 7.5). In Anlehnung an die Arbeit von Carstensen el al. (2012) wurden bei den Fixierungsvarianten die Ankeritems sukzessiv berücksichtigt, d. h., dass zunächst nur die 20 Ankeritems aus der ES K6, nachfolgend die 31 Ankeritems der ES K7 und letztlich alle 51 Ankeritems zur Ermittlung der Kompetenzentwicklung genutzt wurden. Die Differenzierung der Ankeritems ist von besonderem didaktischen Interesse, da sie Auskunft über die Beherrschung unterschiedlicher Lerninhalte gibt. Diesbezüglich ist zu beachten, dass die Skalierungsvarianten bei allen Verfahren angewendet werden konnten, aber die Fixierungsvarianten aufgrund der automatischen Fixierung der Itemschwierigkeiten nur bei der getrennten Skalierung und der Skalierung mit latenten Dimensionen angewendet werden konnten. Abschließend erfolgte eine Gegenüberstellung der drei Verfahren zur längsschnittlichen Kompetenzmodellierung. Dabei wurde insbesondere auf die unterschiedlichen Skalierungs- und Fixierungsvarianten eingegangen, um Schlussfolgerungen für die Modellierung der Rechtschreibkompetenz im Längsschnitt abzuleiten. Die Ergebnisse finden sich in Kapitel 8.2.

F3: Verändert sich die Kompetenzstruktur von Klassenstufe 6 bis 7, und wenn ja, in welcher Weise?

Um diese Fragen zu beantworten, wurden mehrdimensionale längsschnittliche Skalierungen mit den Daten der Entwicklungsstudien (ES) in Klassenstufe 6 (K6) und 7 (K7) mit 307 Schülerinnen und Schülern vorgenommen, um das differenzielle Rechtschreibkompetenzmodell auf eventuelle Veränderungen hin vergleichend zu untersuchen. Dazu wurde auf das zur Beantwortung der zweiten Forschungsfrage eingesetzte methodische Vorgehen und die dort ge-

wonnenen Ergebnisse (vgl. Kapitel 8.2) zurückgegriffen: Die Entwicklung der Teilkompetenzen wurde durch getrennte Skalierungen mit latenten Dimensionen modelliert und die Kompetenzstruktur wurde beschrieben. Zur Klärung der Frage, ob sich die Kompetenzstruktur verändert hat, wurden die ermittelten Personen- und Modellparameter nähergehend analysiert. Anhand der Personenparameter in Form der geschätzten Personenfähigkeit der Schülerinnen und Schüler wurde die ermittelte Kompetenzstruktur im Hinblick auf die für diese Arbeit zentralen Einflussfaktoren tiefergehend differenziert, um auf deskriptiver Ebene unterschiedliche Kompetenzstrukturen zu identifizieren. Fehlende Werte wurden bei den Einflussfaktoren durch einen listenweisen Fallausschluss (Listwise-Deletion) von den Analysen ausgeschlossen, wodurch sich die Stichprobengröße je nach Faktor verringerte. Es wurden die in Tabelle 7.1 beschriebenen Einflussfaktoren mit den jeweiligen Ausprägungen berücksichtigt, die aus den Daten der Schüler- und Elternbefragungen gewonnen wurden. Ebenso sind für die Analysen notwendige Rekodierungen angegeben.

Tabelle 7.1: Kontrollierte Einflussfaktoren

Variable	Ausprägung	Rekodierung
Geschlecht	0 = weiblich 1 = männlich	
Alter der Schülerinnen und Schüler in Jahren[12]	offene Antwortkategorie	
Migrationshintergrund[13]	1 = beide Elternteile 2 = ein Elternteil 3 = kein Elternteil	0 = mindestens ein Elternteil 1 = kein Elternteil
Sprache im Haushalt	0 = andere Sprache 1 = Deutsch	
Anzahl Bücher im Haushalt	1 = keine oder nur sehr wenige (0-10 Bücher) 2 = genug, um ein Regalbrett zu füllen (11-25 Bücher) 3 = genug, um mehrere Regalbretter zu füllen (26-100 Bücher) 4 = genug, um ein kleines Regal zu füllen (101-200 Bücher) 5 = genug, um ein großes Regal zu füllen (201-500 Bücher) 6 = genug, um eine Regalwand zu füllen (mehr als 500 Bücher)	0 = 100 oder weniger Bücher 1 = 101 und mehr Bücher
Schulform	1 = Hauptschule 2 = Schulen mit mehreren Bildungsgängen (SMB) 3 = Realschule 4 = Gesamtschule 5 = Gymnasium	0 = andere Schulform 1 = Gymnasium

Abschließend wurde auf der Ebene der Modellparameter die Dimensionalität des Kompetenzkonstrukts fokussiert und im Längsschnitt mögliche Änderungen der latenten Korrelationen und Reliabilitäten untersucht sowie die unterschiedlichen Modellgeltungstests längsschnittlich bewertet. Da im Hinblick auf die Analyse der Rechtschreibkompetenzstruktur im Längs-

[12] Variable aus Geburtsdatum gebildet.
[13] Variable mithilfe des Geburtslandes der Eltern gebildet.

schnitt ein hoher Forschungsbedarf besteht, wurde das methodische Vorgehen explorativ angelegt. Die Analysen und Befunde sind in Kapitel 8.3 dargestellt.

F4: Beeinflussen die Entwicklung in den Teilkompetenzen bzw. die Faktoren Geschlecht, Alter, sozioökonomischer Hintergrund, Migrationshintergrund und Schulform die Entwicklung der Rechtschreibkompetenz auf der Ebene des ganzen Wortes von Klassenstufe 6 und 7, und wenn ja, in welcher Weise?

Die hierfür eingesetzten Analysen mit der Statistiksoftware SPSS 22 beruhen auf Produkt-Moment-Korrelationen und multiplen Regressionsanalysen (Kapitel 3.4), um Teilkompetenzen des differenziellen Kompetenzkonstrukts und Einflussfaktoren korrelations- und regressionsanalytisch zu identifizieren, die einen prädiktiven Beitrag zur Kompetenzentwicklung auf der Ebene des ganzen Wortes leisten. Da sich Differenzen zweier Kompetenzwerte besonders für die Erklärung von Kompetenzentwicklungen durch Einflussfaktoren eignen, wurden für das ganze Wort und die fünf Teilkompetenzen jeweils die Differenz der Kompetenzwerte aus den beiden Messzeitpunkten der Klassenstufen 6 und 7 zugrunde gelegt (vgl. Hartig, Jude & Wagner, 2008, S. 52). Hierzu wurde als abhängige bzw. zu erklärende Variable (AV) die Differenz der Kompetenzwerte für das ganze Wort aus den beiden Messzeitpunkten der Klassenstufen 6 und 7 definiert. Die unabhängigen bzw. erklärenden Variablen (UV) resultieren aus den Differenzenwerten für die Personenfähigkeiten der fünf Teilkompetenzen und den möglichen Einflussfaktoren, die in dieser Arbeit untersucht werden (vgl. Tabelle 7.1). Fehlende Werte wurden per listenweisen Fallausschluss von den Analysen ausgeschlossen. Konkret wurden in einem ersten Schritt die bivariaten korrelativen Zusammenhänge der Entwicklungen der jeweiligen Teilkompetenzen in Bezug auf die Kompetenzentwicklung beim ganzen Wort bestimmt. Weitergehend wurden die Einflüsse aller fünf Teilkompetenzen bezüglich der Entwicklung auf der Ebene des ganzen Wortes betrachtet, indem eine multiple Regressionsanalyse vorgenommen wurde. Dieses Vorgehen wurde in einem zweiten Schritt analog für die in dieser Arbeit kontrollierten Einflussfaktoren in Bezug auf das ganze Wort wiederholt. In einem letzten Schritt fand die gemeinsame regressionsanalytische Untersuchung der abhängigen Variable in Form der zu erklärenden Kompetenzdifferenz beim ganzen Wort durch die fünf Teilkompetenzen und die ausgewählten Einflussfaktoren als unabhängige Variablen statt. Mit diesem Vorgehen konnte schrittweise und detailliert die Wirkung von Einflussfaktoren auf die Kompetenzent-

wicklung untersucht werden, indem zunächst bivariate und später multiple Analysen vorgenommen wurden. Vorgehensweise und Befunde sind Kapitel 8.4 zu entnehmen.

Am Ende dieses Kapitels wird das Forschungsvorhaben in Tabelle 7.2 in einer Übersicht zusammengestellt.

Tabelle 7.2: Forschungsvorhaben im Überblick

Forschungsfragen (Kapitel 7.1)	**Datengrundlage** (Kapitel 7.2)	**Analysemethoden** (Kapitel 3)	**Ergebnisse** (Kapitel 8)
F1: Wie verhalten sich die in der Entwicklungsstudie und in der Haupterhebung in Klassenstufe 5 ermittelten Gütekriterien für den Rechtschreibtest zueinander?	Entwicklungsstudie in Klassenstufe 5 (N = 298) Haupterhebung in Klassenstufe 5 (N = 4989)	Datenauswertung der Entwicklungsstudie und Haupterhebung in Klassenstufe 5 Vergleich der Entwicklungsstudie mit der Haupterhebung in Klassenstufe 5	Kapitel 8.1
F2: Mit welchen methodischen Verfahren lässt sich die Entwicklung der Rechtschreibkompetenz verlässlich und effizient erfassen?	Längsschnitt: Entwicklungsstudie in Klassenstufe 6 (N = 307) Entwicklungsstudie in Klassenstufe 7 (N = 307)	Vergleich der längsschnittlichen Kompetenzmodellierung auf Ebene des ganzen Wortes: 1. Getrennte Skalierung 2. Skalierung mit virtuellen Personen 3. Skalierung mit latenten Dimensionen	Kapitel 8.2
F3: Verändert sich die Kompetenzstruktur von Klassenstufe 6 bis 7, und wenn ja, in welcher Weise?	Längsschnitt: Entwicklungsstudie in Klassenstufe 6 (N = 307) Entwicklungsstudie in Klassenstufe 7 (N = 307)	Modellierung der Kompetenzstruktur auf Ebene der Teilkompetenzen	Kapitel 8.3
F4: Beeinflussen die Entwicklung in den Teilkompetenzen bzw. die Faktoren Geschlecht, Alter, sozioökonomischer Hintergrund, Migrationshintergrund und Schulform die Entwicklung der Rechtschreibkompetenz auf der Ebene des ganzen Wortes von Klassenstufe 6 und 7, und wenn ja, in welcher Weise?	Längsschnitt: Entwicklungsstudie in Klassenstufe 6 (N = 307) Entwicklungsstudie in Klassenstufe 7 (N = 307)	Korrelations- und Regressionsanalysen	Kapitel 8.4

7.2 DATENGRUNDLAGE

Die kleineren Entwicklungsstudien in den Klassenstufen 5 bis 7 (ES K5-K7) basieren auf einer nicht zufällig gezogenen Stichprobe zur Erprobung der Fragebögen und Testinstrumente, wobei freiwillig gemeldete Schulen aller Schulformen in den Jahren 2009 (K5) und 2011 (K6 und K7) ausgewählt wurden.

Die Stichprobe der Haupterhebung in Klassenstufe 5 (HE K5) der Sekundarstufe I wurde zufällig im Jahr 2011 gezogen.[14] Auf institutioneller Ebene erfolgte dafür eine mehrstufige Zufallsauswahl, d. h., dass zunächst eine Auswahl aus allen Regelschulen getroffen wurde, um nachfolgend geeignete Schülerinnen und Schüler auszuwählen (vgl. Aßmann et al., 2011, S. 54).

Tabelle 7.3 liefert eine Übersicht zur Beschreibung der Stichprobe anhand der Merkmale „Schulform", „Schüler" und „Geschlecht" unter Angabe der absoluten und relativen Häufigkeit. In Bezug auf den Längsschnitt zwischen den Klassenstufen 6 und 7 ist anzumerken, dass nicht auf die gesamte Stichprobe zurückgegriffen wird, sondern nur die Schülerinnen und Schüler berücksichtigt werden, die an beiden Testzeitpunkten teilgenommen haben. Dementsprechend verbleiben 307 Schülerinnen und Schüler der 414 Teilnehmer in der Klassenstufe 6, was einem Anteil von ungefähr 74 Prozent entspricht.

[14] Diese Arbeit nutzt Daten des Nationalen Bildungspanels (NEPS) Startkohorte 3 (Klasse 5), doi:10.5157 / NEPS:SC3:1.0.0.

Tabelle 7.3: Ausgewählte Merkmale der Stichproben

		ES K5		ES K6 | ES K7		HE K5	
		Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
Schulform	Grundschule[15]	-	-	-	-	14	6,90
	Hauptschule	3	20,00	3	17,65	42	20,69
	SMB	2	13,33	4	23,53	24	11,82
	Realschule	3	20,00	1	5,88	38	18,72
	Gesamtschule	1	6,67	1	5,88	10	4,93
	Gymnasium	6	40,00	8	47,06	75	36,95
	Gesamt	**15**	**100,00**	**17**	**100,00**	**203**	**100,00**
Schüler	Grundschule	-	-	-	-	274	5,49
	Hauptschule	45	15,10	40	13,03	637	12,77
	SMB	33	11,07	53	17,26	480	9,62
	Realschule	54	18,12	22	7,17	1.085	21,75
	Gesamtschule	19	6,38	19	6,19	249	4,99
	Gymnasium	147	49,33	173	56,35	2.264	45,38
	Gesamt	**298**	**100,00**	**307**	**100,00**	**4.989**	**100,00**
Geschlecht	Jungen	154	51,68	164	53,42	2.407	48,27
	Mädchen	144	48,32	142	46,25	2.580	51,73
	Gesamt	**298**	**100,00**	**306**[16]	**99,67**	**4.987**[17]	**100,00**

Die Datenbasis der Entwicklungsstudie in Klassenstufe 5 (ES K5) umfasst 15 Schulen und 298 Schülerinnen und Schüler, die in Bezug auf die fünf Schulformen eine ungleichmäßige Verteilung aufweist. Mit jeweils 20 Prozent sind die Haupt- und Realschulen in der Stichprobe vertreten. Unterrepräsentiert sind dagegen die Schulen mit mehreren Bildungsgängen (SMB) mit einem Anteil von 13,33 Prozent und die Gesamtschulen mit einem geringen Anteil von 6,67 Prozent. Dahingegen sind die Gymnasien überrepräsentiert und weisen einen Anteil von 40 Prozent auf. Dementsprechend wurden an sechs Gymnasien insgesamt 147 Schülerinnen und Schüler getestet. An jeweils drei Schulen wurden 45 bzw. 54 Schülerinnen und Schüler der Haupt- und Realschulen getestet. Aus den zwei SMB und der einen Gesamtschule stehen weitere 33 bzw. 19 Teilnehmer der Studie zur Verfügung. Von den 289 Schülerinnen und Schülern aus der 5. Klassenstufe waren 154 Jungen (51,68 %) und 144 Mädchen (48,32 %).

[15] Die Haupterhebung in Klasse 5 fand auch in Berlin und Brandenburg statt, wo es eine sechsjährige Grundschulzeit gibt und sich folglich die getesteten Schülerinnen und Schüler noch in der Grundschule befanden.
[16] Die Angabe für das Geschlecht liegt für einen der 307 Schülerinnen und Schülern nicht vor.
[17] Die Angabe für das Geschlecht liegt für zwei der 4.989 Schülerinnen und Schülern nicht vor.

Aus der Entwicklungsstudie in Klassenstufe 6 und 7 liegen Daten aus 17 Schulen von 307 Schülerinnen und Schülern im Längsschnitt vor. Die Stichprobe basiert auf einer ungleichmäßigen Verteilung der Stichproben, wobei die Realschulen und Gesamtschulen mit lediglich 5,88 Prozent unterrepräsentiert und die Gymnasien erneut mit 47,06 Prozent überrepräsentiert sind. Mit 23,53 bzw. 17,65 Prozent sind die Schulen mit mehreren Bildungsgängen und Hauptschulen vertreten. Insgesamt 41 Schülerinnen und Schüler wurden an jeweils einer Realschule (22) und einer Gesamtschule (19) getestet. Aus den drei bzw. vier Hauptschulen (40) und Schulen mit mehreren Bildungsgängen (53) stehen Daten von insgesamt 93 Schülerinnen und Schülern zur Verfügung. Der größte Anteil der Schülerinnen und Schüler stammt aus den acht Gymnasien mit insgesamt 173 Teilnehmern. In Bezug auf das Geschlecht liegt ein etwas größerer Anteil an Jungen mit 54,07 Prozent im Vergleich zu 45,93 Prozent Mädchen vor.

Die Datenbasis der Haupterhebung in Klassenstufe 5 (HE K5) umfasst 203 Schulen und 4.989 Schülerinnen und Schüler. In Bezug auf die fünf Schulformen ist wie bei der Entwicklungsstudie in Klassenstufe 5 eine ungleichmäßige Verteilung zu erkennen. Eine Besonderheit der Stichprobe ist die Teilnahme von 14 Grundschulen, die einem Anteil von 6,90 Prozent entsprechen.[18] Mit 20,69 Prozent sind die Hauptschulen mit insgesamt 42 Schulen in der Stichprobe vertreten. Die Schulen mit mehreren Bildungsgängen (SMB) haben einen Anteil von 11,72 Prozent mit insgesamt 24 Schulen. Mit 18,72 Prozent und 38 Schulen sind etwas mehr Realschulen in der Stichprobe vertreten. Die kleinste Gruppe machen die Gesamtschulen mit zehn Schulen aus, die lediglich einen Anteil von 4,93 Prozent aufweisen. Dagegen sind die Gymnasien die stärkste Gruppe mit 36,95 Prozent und 75 Schulen. An den Grundschulen nahmen 274 Schülerinnen und Schüler an der Erhebung teil. 637 bzw. 480 Schülerinnen und Schüler wurden an den Hauptschulen und SMB und 1.085 Schülerinnen und Schüler an den Realschulen befragt und getestet. Aus den Gesamtschulen nahmen 249 Schülerinnen und Schüler teil. Mit 2.264 Schülerinnen und Schülern sind die Gymnasien überproportional vertreten. In Bezug auf das Geschlecht zeigt sich eine nahezu gleiche Verteilung, da Daten von 2.407 Jungen (48,27 %) und 2.580 Mädchen (51,73 %) vorliegen.

Im nächsten Abschnitt wird auf das Testinstrument zur kompetenzorientierten Leistungsmessung der Rechtschreibung im Rahmen von NEPS eingegangen.

[18] Die Teilnahme von Grundschulen ist durch die sechsjährige Grundschule in zwei Bundesländern bedingt.

7.3 TESTINSTRUMENT

Der sprachsystematische Rechtschreibtest in NEPS (Kapitel 5.2) ist als kombinierter Lücken- und Satztest konzipiert, wobei der Anteil der Sätze pro Klassenstufe ansteigt. Dies hängt mit den neuen Testinhalten zusammen, die den jeweiligen Lerninhalten entsprechen. Das längsschnittliche Testdesign enthält zu jedem Messzeitpunkt identische Ankeritems, die eine Verbindung bzw. Verankerung zwischen den einzelnen Erhebungen herstellen. So ist es möglich, mithilfe der Ankeritems die Entwicklung der Personenfähigkeit zu ermitteln (vgl. Lord, 1980; Hambleton, Swaminathan & Rogers, 1991). Das Testinstrument wurde für jede Klassenstufe auf Grundlage der Auswertungsergebnisse weiterentwickelt. Tabelle 7.4 liefert einen Überblick über die Anzahl der Wörter und Struktureinheiten als Analyseitems in den Tests für die Klassenstufen 5 bis 7.

Tabelle 7.4: Das Testinstrument von Klassestufe 5–7

	Wörter	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
		Struktureinheiten				
ES K5	54 (32 Lücken/ 3 Sätze)	51	49	24	71	38
ES K6	75 (22 Lücken/ 6 Sätze)	49	50	39	89	72
ES K7	92 (22 Lücken/ 9 Sätze)	61	76	45	111	91
HE K5	54 (32 Lücken/ 3 Sätze)	51	49	24	71	38

ES = Entwicklungsstudie
HE = Haupterhebung

Die Schülerinnen und Schüler der Entwicklungsstudie (ES) und Haupterhebung (HE) in Klassenstufe 5 (K5) haben 54 ganze Wörter zu verschriftlichen, aus denen 233 Struktureinheiten für die differenzielle Analyse der Rechtschreibkompetenz abgeleitet wurden. Für die Weiterentwicklung des Testinstruments für die Klassenstufe 6 wurden Wörter als Ankeritems ausgewählt, die als ganze Wörter und/oder in Bezug auf darin enthaltene Struktureinheiten über

eine hohe Itemschwierigkeit verfügen. Im Detail verteilen sich die Ankeritems je Messzeitpunkt wie in Tabelle 7.5 zusammengefasst.

Tabelle 7.5: Ankeritems und zeitpunktspezifische Items je Messzeitpunkt

	K5		K6		K7	
	GW	SE	GW	SE	GW	SE
Ankeritems zu K5, K6 & K7	20	93	20	93	20	93
Ankeritems zu K5 & K6	7	26	7	26		
Ankeritems zu K6 & K7			31	133	31	133
Zeitpunktspezifisch zu K5	27	116				
Zeitpunktspezifisch zu K6			17	50		
Zeitpunktspezifisch zu K7					41	161
Gesamtzahl	**54**	**235**	**75**	**302**	**92**	**387**

GW = Ganzes Wort | SE= Struktureinheiten

Ausgehend von dem Testinstrument in Klassenstufe 5 wurden für den Test in Klassenstufe 6 27 Wörter (36 %) mit 119 Struktureinheiten (39,40 %) als Ankeritems ausgewählt. 48 Wörter (64 %) mit 183 Struktureinheiten (60,60 %) kamen hinzu, die Lerninhalte in der 6. Klassenstufe darstellen. Somit bestand das Testinstrument in Klassenstufe 6 aus 75 Wörtern mit 302 Struktureinheiten. Für den Test in Klassenstufe 7 wurden 20 Wörter (21,74 %) mit 93 Struktureinheiten (24,03 %) aus Klassenstufe 5 und 31 Wörter (33,70 %) mit 133 Struktureinheiten (34,37 %) aus Klassenstufe 6 als Ankeritems ausgewählt. 41 Wörter (44,57 %) mit 161 Struktureinheiten (41,60 %) kamen den erweiterten Lerninhalten für Klassenstufe 7 entsprechend hinzu. Der sprachsystematische Rechtschreibtest in Klassenstufe 7 umfasst folglich 92 Wörter mit 387 Struktureinheiten.

Bei der Durchführung des sprachsystematischen Rechtschreibtests in den NEPS-Erhebungen wurde die Objektivität dadurch gewährleistet, dass Test und Testanweisung von einer professionellen Sprecherin auf CD aufgenommen und abgespielt wurden, wodurch die Testabläufe unabhängig vom jeweiligen Testleiter gleichblieben.

Die mit dem Testinstrument gewonnenen Testdaten wurden vom Data Processing and Research Center (DPC) in Hamburg transkribiert und mit einem eigens entwickeltem SRT-Editor

(Frahm, 2013) kodiert. In der ersten Version des SRT-Editors werden nicht geschriebene Wörter als Fehler und nicht als sogenannte Missings kodiert.[19] Dies wurde inzwischen behoben und im Herbst 2014 wird ein entsprechendes SUF-Update für die Haupterhebung K5 erscheinen. Die Analyse der kodierten Daten erfolgte mit der Statistiksoftware SPSS 22. Die Skalierungssoftware ConQuest 2.0 wurde zur Schätzung der Item- und Personenparameter verwendet. Die so geschätzte Personenfähigkeit wurde in Anlehnung an die Standards der Large-Scale-Assesments auf einen Mittelwert von 500 und eine Standardabweichung von 100 normiert. Vorab wurden Konstanten und besonders häufig gelöste Items von den Analysen ausgeschlossen.

[19] In Rechtschreibtests sind Fehler und Missings nicht eindeutig voneinander abzugrenzen. Um eine Entscheidungsgrundlage zu gewinnen, wurde der Umfang der Missings mit den Daten der HE K5 anhand der digital erfassten Schülerschreibungen ermittelt. Dieser liegt bei 1,45 Prozent (4.275 Missings von 294.351 Schreibungen) (Blatt & Prosch 2013: unveröffentlichte Quelle). Dies führte dazu, dass die Kodiersoftware um die Funktion erweitert wird, fehlende ganze Wörter als Missings auszuweisen. Fehlende Satzzeichen werden jedoch als Fehler gezählt.

8. ERGEBNISSE DER QUANTITATIVEN DATENANALYSE

Diese Arbeit verfolgt das Ziel, Daten aus NEPS-Studien zur Domäne der Rechtschreibung auszuwerten. Leitend sind hierbei die aufgestellten Forschungsfragen in Kapitel 7.1, die gleichzeitig die Analyseziele festlegen. Zunächst werden die Ergebnisse zur Testentwicklung für die Entwicklungsstudie (ES) und Haupterhebung (HE) in Klassenstufe 5 (K5) dargestellt und anschließend verglichen. Im Hinblick auf die Entwicklung der Personenfähigkeit und der Kompetenzstruktur werden nachfolgend die beiden Entwicklungsstudien (ES) in Klassenstufe 6 (K6) und 7 (K7) herangezogen. Weiterhin werden Befunde der Korrelations- und Regressionsanalysen zur Überprüfung von Einflussfaktoren vorgestellt. Die Ergebnisse zu jeder Forschungsfrage werden abschließend zusammengefasst und interpretiert.

8.1 ERGEBNISSE ZUR TESTENTWICKLUNG

In diesem Kapitel werden die Analyseergebnisse dargestellt, die zur Beantwortung der ersten Forschungsfrage durchgeführt wurden:

F1: Wie verhalten sich die in der Entwicklungsstudie und in der Haupterhebung in Klassenstufe 5 ermittelten Gütekriterien für den Rechtschreibtest zueinander?

Es handelt sich um

- die Überprüfung der Normalverteilung für die Personenparameter, um eine mögliche Verzerrung der Daten festzustellen. Weiterhin wurde ein grafischer Modelltest zur Annahme des Rasch-Modells durchgeführt.
- den Vergleich von ein- und mehrdimensionalen Skalierungsmodellen für das ganze Wort und die Teilkompetenzen, bei dem mögliche Veränderungen aufgrund von statistisch auffälligen Items untersucht werden. Dazu werden die Ursprungs- und die optimierten Ausgangsmodelle gegenübergestellt. Dies zielt darauf ab, geeignete Items für die längsschnittliche Testentwicklung zu identifizieren.

In Kapitel 8.1.1 werden die Befunde für die Entwicklungsstudie in Klassenstufe 5 und in Kapitel 8.1.2 für die Haupterhebung in Klassenstufe 5 vorgestellt und erläutert. In Kapitel 8.1.3 folgt eine Gegenüberstellung und Interpretation der Ergebnisse mit dem Ziel, die Forschungsfrage zu beantworten.

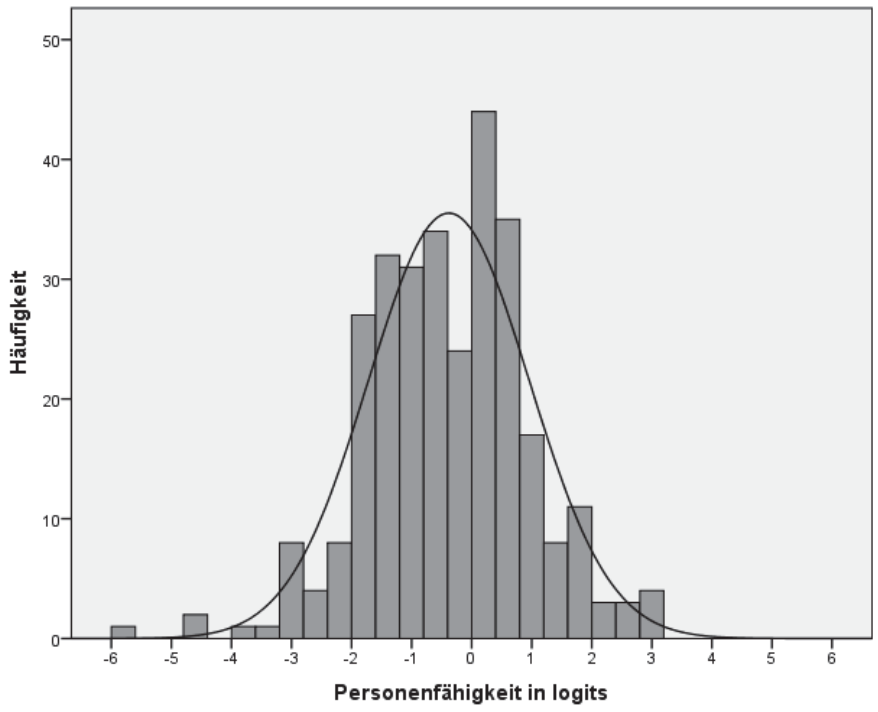
8.1.1 DATENGRUNDLAGE DER ENTWICKLUNGSSTUDIE K5

Für die Daten der Entwicklungsstudie in Klassenstufe 5 (ES K5) wird zunächst die Datenprüfung auf Ebene des ganzen Wortes veranschaulicht. Diese umfasst die Überprüfung einer Normalverteilung und den grafischen Modelltest zur Annahme des Rasch-Modells. Daran knüpft die Ergebnisdarstellung der ein- und mehrdimensionalen Skalierungen auf Grundlage der Item-, Personen- und Modellparameter des ganzen Wortes und der Teilkompetenzen an. Abschließend wird die Dimensionalität des Kompetenzkonstrukts aufgezeigt.

8.1.1.1 DATENPRÜFUNG

Zunächst wird zur Bestimmung der Ausgangslage auf die Überprüfung einer Normalverteilung für die Personenparameter im Fall des ganzen Wortes für die ES K5 auf die Ergebnisse der Kolmogorov-Smirnov-Tests eingegangen. Abbildung 8.1 stellt dazu in einem Histogramm die geschätzten Personenfähigkeiten in Logits gemeinsam mit der Normalverteilungskurve grafisch dar.

Abbildung 8.1: Verteilung der Personenfähigkeit für das ganze Wort – ES K5 (N = 298)

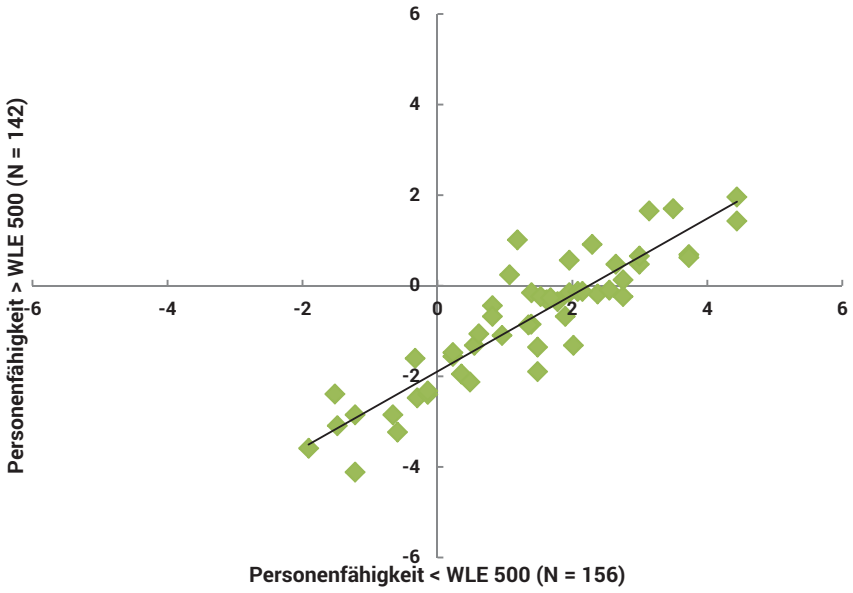


Die Stichprobe der ES K5 weist für die Personenfähigkeit einen Mittelwert von -0.38 Logits und eine Standardabweichung von 1.34 Logits auf. Mithilfe der statistischen Überprüfung kann die Normalverteilung des Personenparameters mit einem Signifikanzniveau von 0.41 nachgewiesen werden, d. h., dass die getesteten Schülerinnen und Schüler die gesamte Fähigkeitsbandbreite abdecken und dementsprechend weniger gute, mittlere und gute Schülerinnen und Schüler an der Testung teilgenommen haben.

Auf die Prüfung einer Normalverteilung für die Teilkompetenzen wird in der Zusammenfassung (Kapitel 8.1.1.2) eingegangen.

Der grafische Modelltest zur Annahme des Rasch-Modells mit 51 Itemschwierigkeiten[20] ist für die beiden Teilstichproben (Personenfähigkeit < oder > 500) der ES K5 in der nachfolgenden Abbildung 8.2 veranschaulicht.

Abbildung 8.2: Grafischer Modelltest auf Basis der Itemschwierigkeiten – ES K5



[20] Von den ursprünglich 54 Itemschwierigkeiten müssen drei Konstanten für den grafischen Modelltest ausgeschlossen werden, die sich bei der Gruppe der Schülerinnen und Schüler mit einer Personenfähigkeit unten dem WLE-Wert von 500 ergeben haben.

Es zeigen sich Differenzen bei den Itemschwierigkeiten zwischen den beiden Gruppen für die ES K5, die sich zwischen 0.17 und 3.38 Logits bewegen. Es wird mit den zwei Teilstichproben eine erklärte Varianz (R^2) in Höhe von 95 Prozent erreicht.

Zusammenfassend lässt sich durch die Prüfung der Personenparameter auf eine Normalverteilung festhalten, dass die Datengrundlage in Bezug auf die Personenfähigkeit keiner Verzerrung bzw. einer schiefen Verteilung unterliegt und somit eine gute Voraussetzung für die weitergehenden Analysen und die Weiterentwicklung des Testinstruments im Rahmen dieser Arbeit bereitstellt.

Weitergehend lässt sich aus dem rein deskriptiven grafischen Modelltest ableiten, dass es aufgrund der Differenzen Hinweise für die Verletzung der Personenhomogenität gibt, die allerdings ohne signifikanzanalytische Prüfung bestehen. Dies könnte wiederum die Annahme einer unterschiedlichen Kompetenzstruktur bei Schülerinnen und Schülern mit einer hohen bzw. niedrigen Personenfähigkeit nahelegen, der im Kapitel 8.3 intensiver nachgegangen wird. Insgesamt lässt sich folgern, dass das Rasch-Modell für die ES K5 mit den erkennbaren Differenzen bei den Itemschwierigkeiten vorläufig angenommen werden kann und die Kompetenzentwicklung modelliert werden darf, um tiefergehende Statistiken zur Modellierbarkeit der Rechtschreibkompetenz zu erhalten.

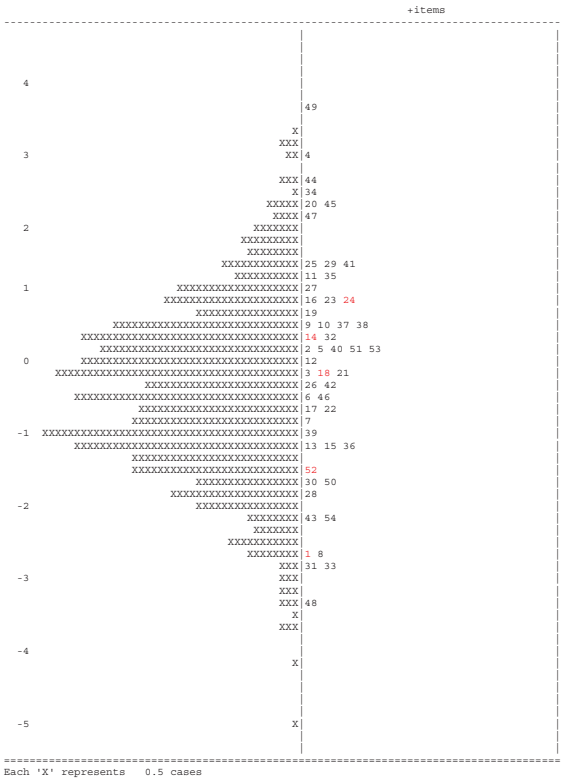
8.1.1.2 EINDIMENSIONALE SKALIERUNGEN

Weitergehend wurde anhand getrennter eindimensionaler Skalierungen mit den Daten aller Wörter bzw. Struktureinheiten für das ganze Wort und die fünf Teilkompetenzen untersucht, wie sich die sogenannten Ursprungsmodelle zu den optimierten Ausgangsmodellen verhalten. Die Ergebnisse in Bezug auf die Item- und Personenparameter sind auf einer gemeinsamen Skala in den Abbildung 8.3 bis Abbildung 8.8 zusammengestellt, die das Verhältnis der Personenfähigkeit auf der linken Seite und der Itemschwierigkeit auf der rechten Seite aufzeigen. Weitergehend werden auf der Ebene der Itemparameter das Verhältnis der leichten und schweren Wörter bzw. Struktureinheiten sowie deren statistische Auffälligkeiten fokussiert. Auf der Ebene der Personenparameter werden die Mittelwerte und Standardfehler der Personenfähigkeit berichtet. Abschließend werden die Reliabilitäten der jeweiligen Skalierungsmodelle als Modellparameter vorgestellt.

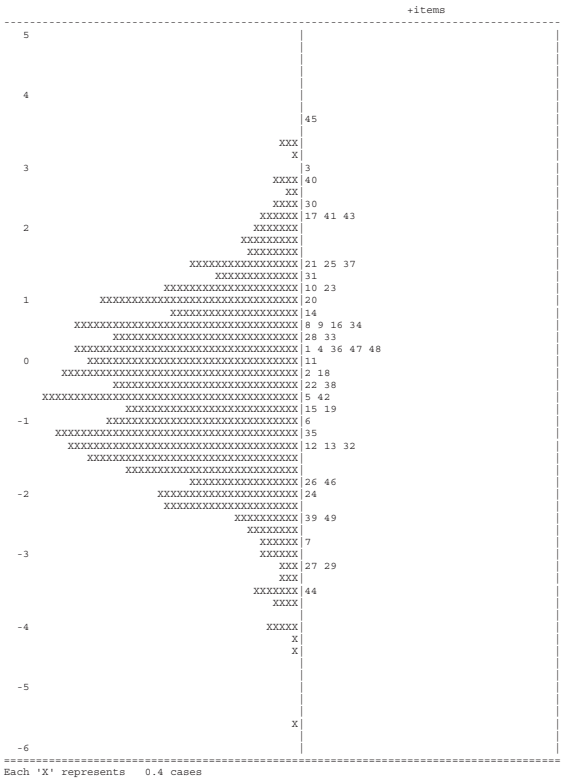
Zunächst werden das Ursprungs- und Ausgangsmodell zum ganzen Wort anhand der Item- und Personenparameter betrachtet.

Abbildung 8.3: Ganzes Wort – ES K5

Ursprungsmodell



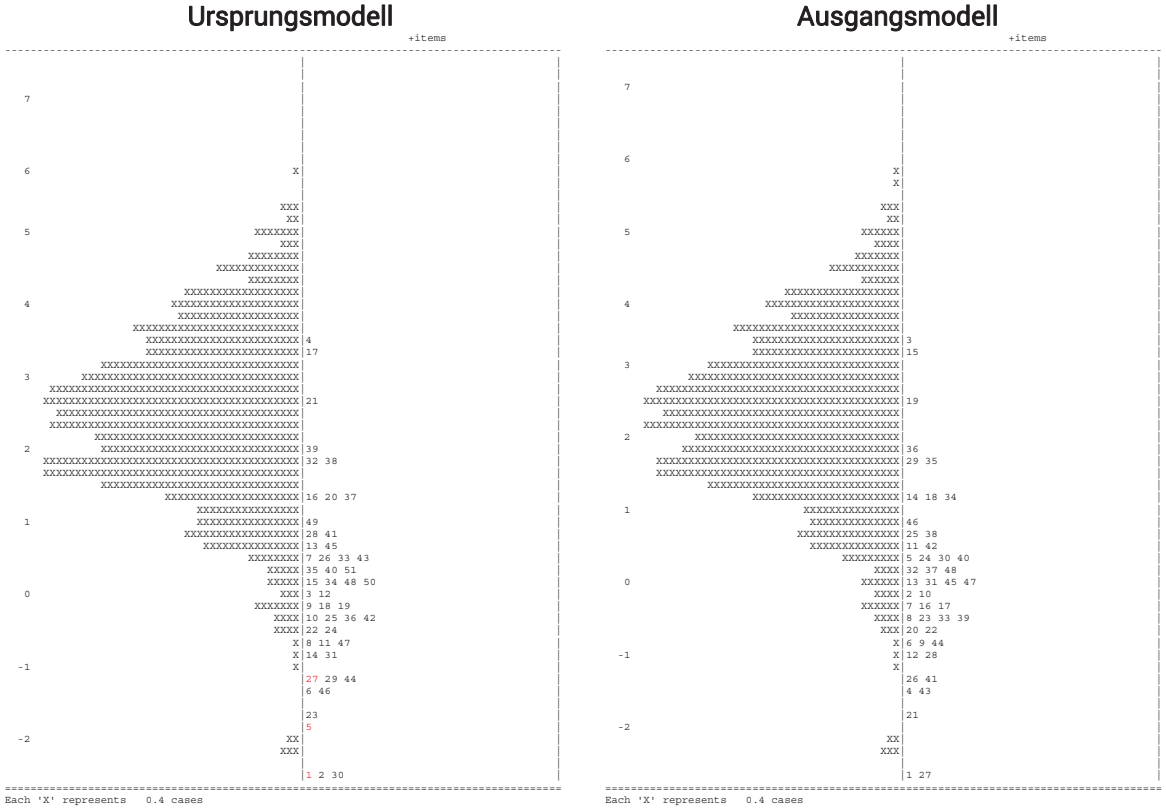
Ausgangsmodell



Das Ursprungsmodell umfasst 54 ganze Wörter, wovon sich 25 (46,30 %) leichte und 29 (56,70 %) schwere Wörter identifizieren lassen. Auffällig sind fünf rot markierte Wörter, die aufgrund von statistischen Kriterien von den Analysen ausgeschlossen werden. Konkret weisen zwei schwere Wörter (14, 24) mit einer Itemschwierigkeit über null Logits einen höheren Item-Fit als 1.20 und eine geringere Trennschärfe als 0.26 auf. Zwei leichte Wörter (18, 52) mit einer Itemschwierigkeit kleiner als null Logits überschreiten die Grenze für den Item-Fit. Ein weiteres leichtes Wort (1) unterschreitet das Kriterium für die Trennschärfe. Die Personenfähigkeit (MW -0.38 | SE 0.08) der Schülerinnen und Schüler ist leicht gemindert. Bei dem resultierenden Ausgangsmodell verbleiben 49 ganze Wörter. Die Personenfähigkeit (MW -0.46 | SE 0.08) verringert sich weiter, da überwiegend leichte Wörter entfernt werden.

Nun folgen die Ergebnisse zu den Teilkompetenzen in einer identischen Darstellungsform, wobei zunächst der Kernbereich, bestehend aus dem phonographisch-silbischen Prinzip sowie dem morphologischen Prinzip, betrachtet werden.

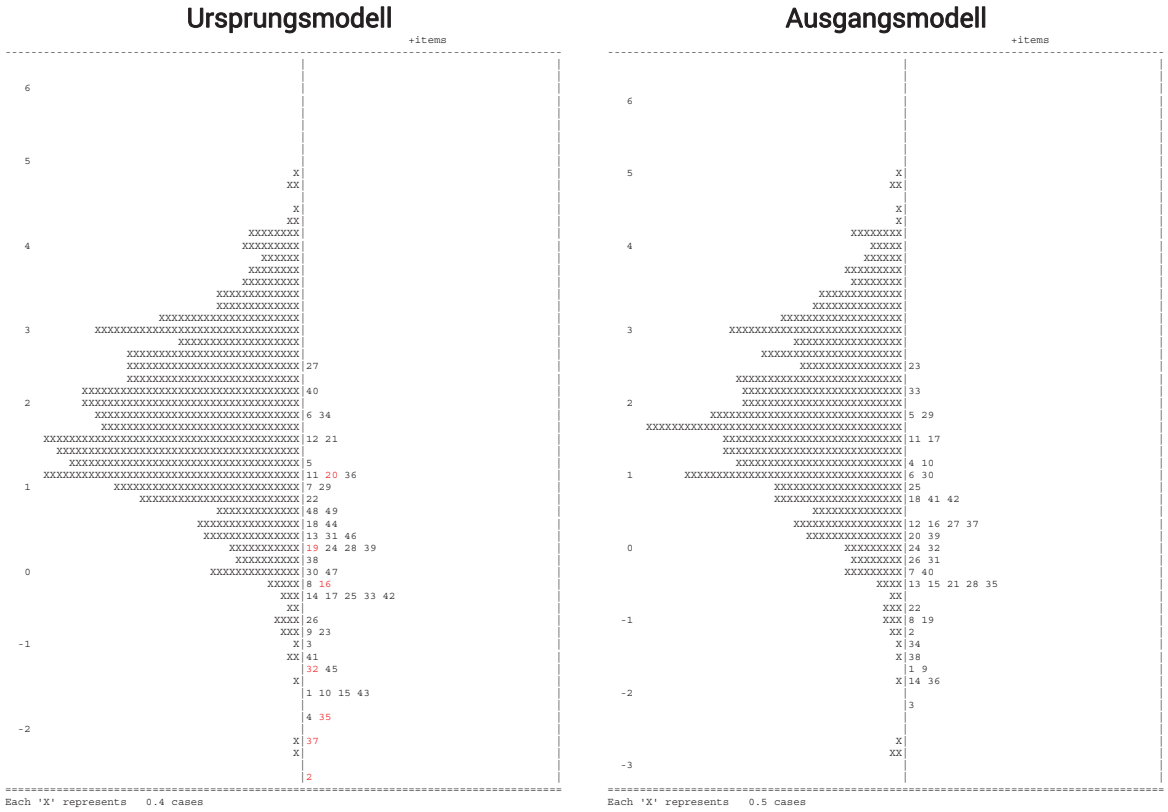
Abbildung 8.4: Phonographisch-silbisches Prinzip im Kernbereich – ES K5



Bei dem phonographisch-silbischen Prinzip liegen 51 Struktureinheiten im Ursprungsmodell vor, die 25 (49,02 %) leichte und 26 (50,98 %) schwere Itemschwierigkeiten aufweisen. Drei leichte Einheiten (1, 5, 27) werden wegen einer zu geringen Trennschärfe von den Analysen ausgeschlossen. Die Personenfähigkeit (MW 2.41 | SE 0.08) deutet auf eine hohe Schülerkompetenz in diesem Prinzip hin. Das Ausgangsmodell mit 48 Struktureinheiten führt zu einer leicht gesunkenen Personenfähigkeit (MW 2.29 | SE 0.08).

Analog erfolgt die deskriptive Darstellung des morphologischen Prinzips des Kernbereichs, wie es in Abbildung 8.5 veranschaulicht ist.

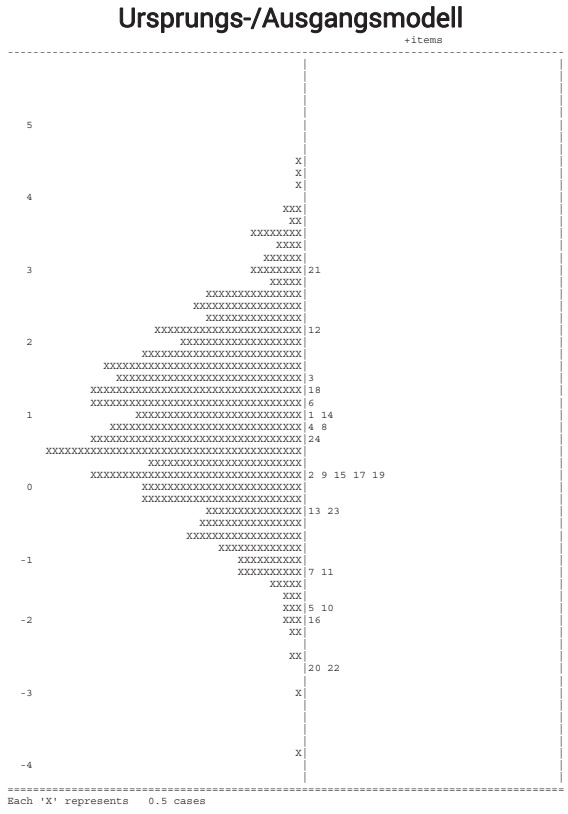
Abbildung 8.5: Morphologisches Prinzip im Kernbereich – ES K5



Das Ursprungsmodell des morphologischen Prinzips umfasst 49 Struktureinheiten, wovon sich 23 (46,94 %) leichte und 26 (53,06 %) schwere Einheiten identifizieren lassen. Es muss eine schwere Einheit (20) aufgrund eines zu hohen Item-Fits sowie einer zu geringen Trennschärfe von den Analysen ausgeschlossen werden. Zudem verfügen fünf leichte Einheiten (2, 16, 32, 35, 37) und eine schwere Einheit (19) über eine zu geringe Trennschärfe. Die Personenfähigkeit (MW 1.80 | SE 0.07) deutet ebenfalls auf eine hohe Kompetenz der Schülerinnen und Schüler bei diesem Prinzip hin. Es verbleiben 43 Struktureinheiten bei dem Ausgangsmodell, wobei eine leicht gesunkene Personenfähigkeit (MW 1.67 | SE 0.07) zu verzeichnen ist.

Weitergehend werden Ergebnisse der deskriptiven Analyse der Item- und Personenparameter zum Peripheriebereich referiert.

Abbildung 8.6: Peripheriebereich – ES K5

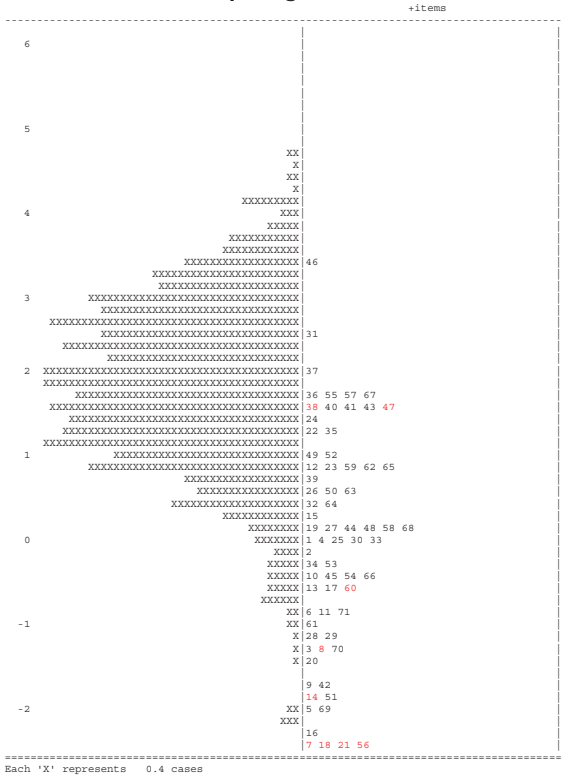


Der Peripheriebereich umfasst im Ursprungsmodell 24 Struktureinheiten, wovon 9 (37,50 %) leichte und 15 (62,50 %) schwere Einheiten vorliegen. Alle Struktureinheiten halten die statistischen Kriterien ein, weshalb es nicht erforderlich ist, einzelne Einheiten auszuschließen. Die Personenfähigkeit (MW 0.86 | SE 0.08) deutet daraufhin, dass die Schülerinnen und Schüler im Peripheriebereich weniger kompetent sind.

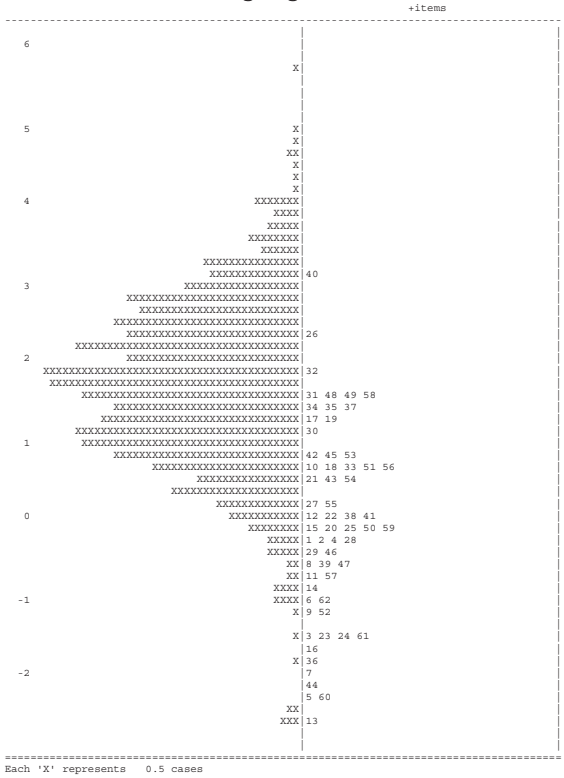
Abschließend werden das Prinzip der Wortbildung und das wortübergreifende Prinzip mithilfe der deskriptiven Analysen für die Item- und Personenparameter dargestellt.

Abbildung 8.7: Prinzip der Wortbildung – ES K5

Ursprungsmodell



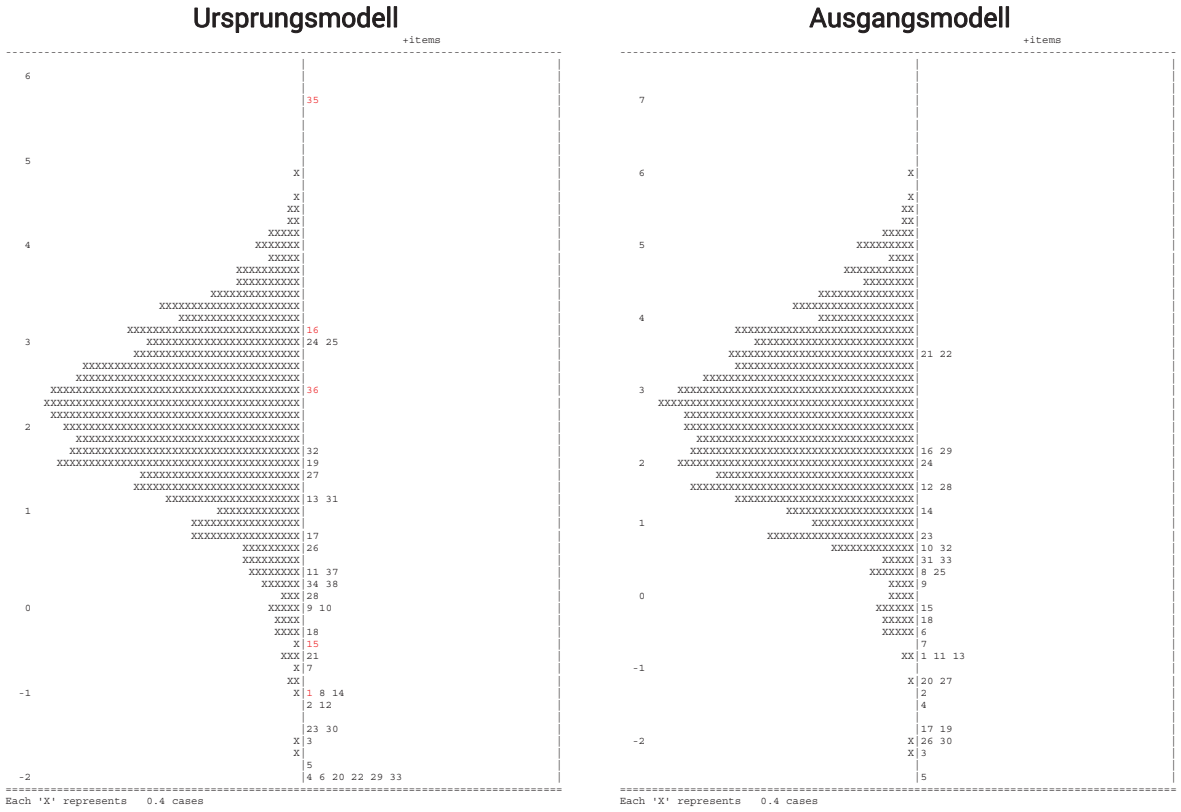
Ausgangsmodell



Das Ursprungsmodell des Prinzips der Wortbildung enthält 71 Struktureinheiten, die sich in 32 (45,07 %) leichte bzw. 39 (54,93 %) schwere Einheiten differenzieren lassen. Es muss eine leichte Einheit (56) wegen der Verletzung beider Kriterien von den Analysen ausgeschlossen werden. Weitere zwei schwere Einheiten (38, 47) weisen einen zu hohen Item-Fit auf und sechs leichte Einheiten (7, 8, 14, 18, 21, 60) verfügen über eine zu geringe Trennschärfe. Auf eine höhere Schülerfähigkeit in dieser Teilkompetenz deutet die Personenfähigkeit (MW 1.68 | SE 0.07) hin. Dies führt zu dem Ausgangsmodell mit insgesamt 62 Struktureinheiten, wobei es zu einer Verringerung der Personenfähigkeit (MW 1.83 | SE 0.07) kommt.

Abschließend erfolgt die deskriptive Darstellung des wortübergreifenden Prinzips in der bekannten Vorgehensweise.

Abbildung 8.8: Wortübergreifendes Prinzip – ES K5



Bei dem wortübergreifenden Prinzip liegen 38 Struktureinheiten vor, wobei sich 20 (52,63 %) leichte bzw. 18 (47,37 %) schwere Einheiten identifizieren lassen. Es müssen fünf Einheiten von den Analysen ausgeschlossen werden. Zwei schwere Einheiten (16, 36) verletzten sowohl das Kriterium für den Item-Fit als auch die Trennschärfe. Zudem verfügen eine schwere (35) und zwei leichte Einheiten (1, 15) über eine zu geringe Trennschärfe. Die Personenfähigkeit (MW 2.05 | SE 0.07) deutet ebenfalls auf eine gut ausgeprägte Kompetenz der Schülerschaft bei diesem Prinzip hin. Das Ausgangsmodell mit den verbleibenden 33 Struktureinheiten weist eine gestiegene Personenfähigkeit (MW 2.49 | SE 0.08) auf.

Inwiefern sich das differenzielle Kompetenzmodell mit dem sprachsystematischen Rechtschreibtest erfassen lässt, wird abschließend geklärt. Dazu sind in Tabelle 8.1 die Reliabilitäten für das ganze Wort und die fünf Teilkompetenzen festgehalten.

Tabelle 8.1: Reliabilitäten – ES K5 – Eindimensionale Skalierungen

	Ganzes Wort	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
	Wörter	Struktureinheiten				
Ursprungsmodell	0.93	0.86	0.87	0.84	0.90	0.79
Ausgangsmodell	0.93	0.85	0.87	-	0.88	0.81

Bereits mit den Ursprungsmodellen ist gewährleistet, dass die Rechtschreibung auf der Ebene des ganzen Wortes und den fünf Teilkompetenzen zuverlässig erfasst wird, da die Reliabilitäten mit Ausnahme des wortübergreifenden Prinzips über 0.80 liegen und somit als zufriedenstellend zu werten sind. Durch den Ausschluss der statistisch auffälligen Wörter bei den Ausgangsmodellen kommt es nur zu sehr geringen Abweichungen bei den Reliabilitäten, die sich zwischen 0.00 und 0.02 bewegen. Demnach ist der sprachsystematische Rechtschreibtest in der Lage, das differenzielle Kompetenzmodell in der Klassenstufe 5 auf eindimensionaler Ebene zu erfassen.

ZUSAMMENFASSUNG DER ERGEBNISSE DER EINDIMENSIONALEN SKALIERUNGEN

Es lässt sich in Tabelle 8.2 festhalten, dass sich von den 54 Testwörtern, die der sprachsystematische Rechtschreibtest in der ES K5 auf der Ebene des ganzen Wortes enthält, lediglich fünf Wörter aus statistischen Gründen nicht für die kompetenzorientierte Leistungsmessung eignen. Auf der Ebene der Teilkompetenzen erfüllen 24 Struktureinheiten nicht die statistischen Kriterien, weshalb von den ursprünglich 233 Struktureinheiten noch 209 Items für die Datenanalyse der eindimensionalen Modelle zur Verfügung stehen.

Tabelle 8.2: Anzahl und statistische Auffälligkeiten der ganzen Wörter und Struktureinheiten – ES K5 – Eindimensionale Skalierungen

	Ganzes Wort	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
	Wörter	Struktureinheiten				
Ursprungsmodell	54	51	49	24	71[21]	38
Item-Fit & Trennschärfe	-2	-	-1	-	-1	-2
Item-Fit	-2	-	-	-	-2	-
Trennschärfe	-1	-3	-6	-	-6	-3
Ausgangsmodell	49	48	42	24	62	33

Insgesamt müssen drei leichte und zwei schwere Wörter sowie 17 leichte und sieben schwere Struktureinheiten von den Analysen ausgeschlossen werden. Damit zeigt sich, dass mehr leichte Wörter und Struktureinheiten die statistischen Kriterien nicht erfüllen. Der aus statistischen Gründen zu rechtfertigende Verzicht auf die statistisch auffälligen Testwörter bzw. Struktureinheiten beläuft sich auf 9.26 Prozent bei dem ganzen Wort und auf 10.30 Prozent bei den Teilkompetenzen. Somit stehen noch ungefähr 90 Prozent der eingesetzten Auswertungseinheiten für die Datenanalyse und Weiterentwicklung des Testinstruments zur Verfügung, das als gute Grundlage zu bezeichnen ist.

Angesichts der ausgewählten deskriptiven Statistiken in Tabelle 8.3 ist bei den Ursprungsmodellen auf der Ebene des ganzen Wortes zu erkennen, dass die Schülerinnen und Schüler über eine etwas verringerte Kompetenz bei der korrekten Bearbeitung der eingesetzten Testwörter verfügen, also dass der sprachsystematische Rechtschreibtest eher etwas schwerer für die Schülerinnen und Schüler ist. Auf der Ebene der Struktureinheiten der Teilkompetenzen ist mit Ausnahme des Peripheriebereichs eine gut ausgeprägte Kompetenz der Schülerschaft in den jeweiligen Prinzipien zu beobachten. Die Normalverteilung der Personenparameter ist bei allen Ursprungsmodellen gegeben.

[21] Eine Struktureinheit wurde von allen Schülerinnen und Schülern richtig gelöst und wurde direkt ausgeschlossen.

Tabelle 8.3: Veränderungen der deskriptiven Statistiken – ES K5 – Eindimensionale Skalierungen

	Ganzes Wort	Phono-graphisch-silbischer Kern-bereich	Morpho-logischer Kern-bereich	Peripherie-bereich	Prinzip der Wort-bildung	Wortüber-greifendes Prinzip
	Wörter	Struktureinheiten				
Ursprungsmodell						
MW	-0.38/0.00	2.41/0.00	1.80/0.00	0.86/0.00	1.83/0.00	2.05/0.00
SE	0.08/0.22	0.08/0.20	0.07/0.17	0.08/0.30	0.07/0.16	0.07/0.30
NV	✓/-	✓/-	✓/-	✓/-	✓/-	✓/-
Ausgangsmodell						
MW	-0.46/0.00	2.29/0.00	1.67/0.00	-	1.68/0.00	2.49/0.00
SE	0.08/0.24	0.08/0.19	0.07/0.18	-	0.07/0.16	0.08/0.28
NV	✓/-	✓/-	✓/-	-	✓/-	×/-

Angabe in Logits für die Personenfähigkeit/Itemschwierigkeit

Durch die Reduzierung der statistisch auffälligen Wörter bzw. Struktureinheiten werden das ganze Wort bzw. die Teilkompetenzen für die Schülerinnen und Schüler schwerer, außer bei dem wortübergreifenden Prinzip, bei dem es zu einer Vereinfachung kommt. Mit Ausnahme des wortübergreifenden Prinzips sind die Personenparameter weiterhin normalverteilt, wobei die Personenfähigkeit eine positive Verzerrung aufweist. Es ist festzuhalten, dass die Schülerinnen und Schüler eher in der Lage sind, die einzelnen Struktureinheiten der Teilkompetenzen korrekt zu schreiben als ein ganzes Wort.

Die aus statistischen Gründen durchgeführte Reduzierung der Wörter bringt weitere Veränderungen mit sich, die nun in Bezug auf die prozentualen Lösungshäufigkeiten für das ganze Wort und die fünf Teilkompetenzen in Tabelle 8.4 aufgezeigt werden.

Tabelle 8.4: Prozentuale Lösungshäufigkeit – ES K5

		Ganzes Wort	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
		Wörter	Struktureinheiten				
Ursprungsmodell	MW	44.26	83.22	77.65	62.99	77.68	78.49
	(SE)	(1.12)	(0.81)	(0.89)	(1.18)	(0.85)	(0.71)
Ausgangsmodell	MW	43.35	82.35	76.08	62.99	76.07	82.48
	(SE)	(1.18)	(0.85)	(0.99)	(1.18)	(0.93)	(0.80)

Angaben in Prozent

Bei den Urspungsmodellen ist auf der Ebene des ganzen Wortes zu erkennen, dass es nur knapp 44 Prozent der Schülerinnen und Schüler gelingt, ein Wort richtig zu schreiben. Wesentlich besser schneiden sie auf der Ebene der Teilkompetenzen ab, wobei die durchschnittliche Lösungshäufigkeit zwischen knapp 63 und 83 Prozent variiert. Besonders häufig werden die Struktureinheiten des phonographisch-silbischen Prinzips im Kernbereich gelöst, gefolgt von dem wortübergreifenden Prinzip. Fast gleich häufig erfolgt die korrekte Bearbeitung der Struktureinheiten des morphologischen Prinzips im Kernbereich und des Prinzips der Wortbildung, die beide morphologische Teilaspekte abdecken. Der Peripheriebereich stellt eine Ausnahme dar, weil diese Teilkompetenz mit Abstand die geringste Lösungswahrscheinlichkeit aufweist. Bei den Ausgangsmodellen ergeben sich Abweichungen, die sich zwischen -1.62 und 3.99 Prozentpunkten bewegen. Bis auf das wortübergreifende Prinzip werden weniger Wörter bzw. Struktureinheiten gelöst. Dies ist dadurch bedingt, dass in den meisten Fällen die als leicht zu bezeichnende Wörter und Struktureinheiten entfernt werden und sich die Personenfähigkeit daraufhin verringert. Im Gegensatz dazu werden bei dem wortübergreifenden Prinzip als schwer zu bezeichnende Wörter aufgrund der ermittelten Gütekriterien entfernt, wodurch der Test für die Schülerinnen und Schüler leichter wird, was sich in einer gestiegenen Personenfähigkeit ausdrückt.

Diese Ergebnisse machen deutlich, dass die Personenparameter in einer logischen Abhängigkeit zu den Itemparametern stehen, d. h., dass sich Änderungen in der Schwierigkeit des Tests unmittelbar auf die Fähigkeit der Schülerinnen und Schüler auswirken. Folglich steigt die Per-

sonenfähigkeit, wenn der Test leichter wird, und die Personenfähigkeit sinkt, wenn der Test schwieriger wird.

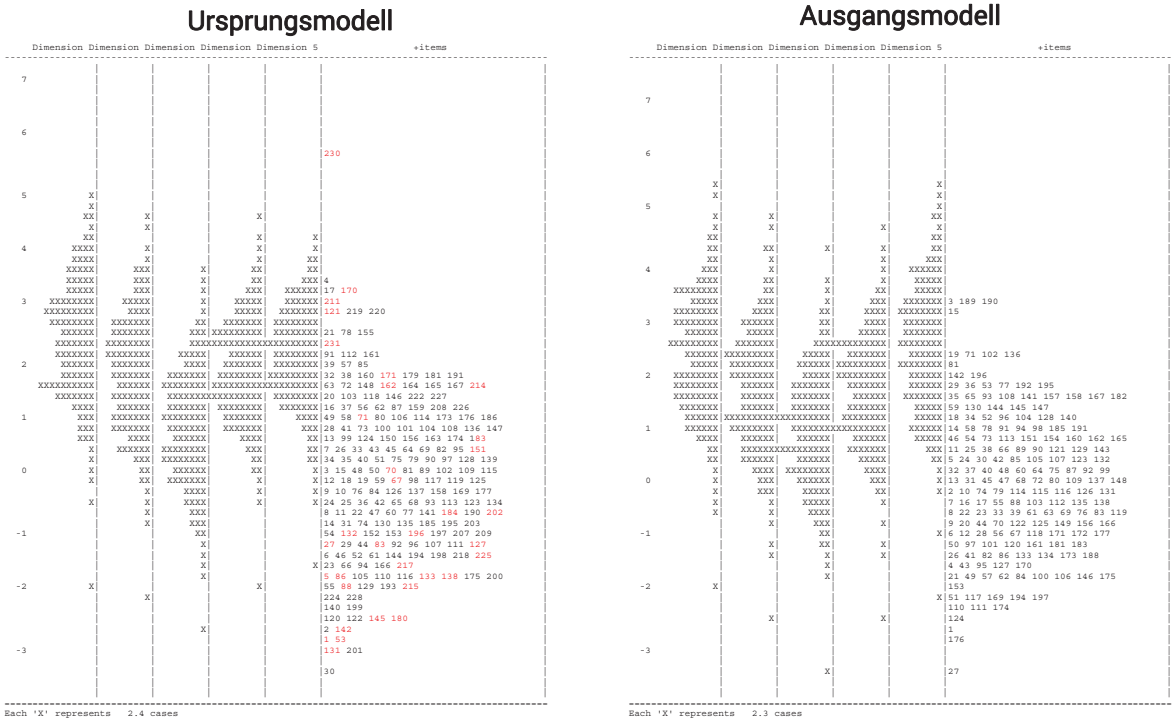
Nach der Ergebnisdarstellung der eindimensionalen Skalierungen folgt nun in analoger Weise die mehrdimensionale Betrachtung der Daten für die ES K5.

8.1.1.3 MEHRDIMENSIONALE SKALIERUNG

Entsprechend der eindimensionalen Betrachtung erfolgt auch für die mehrdimensionale Skalierung zunächst eine umfassende Analyse der Item-, Personen- und Modellparameter zu den fünf Teilkompetenzen, bevor auf die Ergebnisse der Reliabilitäts- und Korrelationsanalyse eingegangen wird.

Dazu werden zunächst die Ergebnisse zu den Item- und Personenparameter der fünfdimensionalen Skalierung für das Ursprungs- und Ausgangsmodell herangezogen, die in Abbildung 8.9 dargestellt sind.

Abbildung 8.9: Mehrdimensionale Skalierung – ES K5



Das Ursprungsmodell der fünfdimensionalen Skalierung umfasst 233 Struktureinheiten, die sich auf die fünf Teilkompetenzen verteilen:

Die ersten beiden Dimensionen auf der linken Seite bilden den Kernbereich, der aus dem phonographisch-silbischen Prinzip und morphologischen Prinzip besteht. Das phonographisch-silbische Prinzip umfasst 51 Struktureinheiten, wovon 25 (49,02 %) leichte und 26 (50,98 %) schwere Einheiten enthalten sind. Bei dem morphologischen Prinzip mit 49 Struktureinheiten lassen sich 23 (46,94 %) leichte und 26 (53,06 %) schwere Einheiten identifizieren. Im Kernbereich erfüllen insgesamt zehn Struktureinheiten nicht die statistischen Kriterien für den Item-Fit (≤ 1.20) und die Trennschärfe (> 0.25), die in der Abbildung rot hervorgehoben sind und von den Analysen ausgeschlossen werden. Konkret unterschreiten drei leichte Einheiten (1, 5, 27) bei dem phonographisch-silbischen Prinzip das Kriterium für die Trennschärfe. Bei dem morphologischen Prinzip überschreiten zwei schwere Einheiten (70, 71) den Grenzwert für den Item-Fit und die Trennschärfe. Weitere fünf leichte Einheiten (53, 67, 83, 86, 88) verfügen über eine zu geringe Trennschärfe. Die hohe Kompetenz der Schülerinnen und Schüler im Kernbereich wird an den Mittelwerten deutlich, die beim phonographisch-silbischen Prinzip (MW 2.40 | SE 0.08) und beim morphologischen Prinzip (MW 1.80 | SE 0.07) auf einem hohen Niveau liegen. Der mittig platzierte Peripheriebereich umfasst 24 Struktureinheiten mit neun (37,50 %) leichten und 15 (62,50 %) schweren Itemschwierigkeiten. Im Gegensatz zur eindimensionalen Skalierung des Peripheriebereichs (Kapitel 8.1.1.2) unterschreitet eine schwere Einheit (121) die Grenze für die Trennschärfe. Die Schülerinnen und Schüler erzielen bei diesem Prinzip die geringste Personenfähigkeit (MW 0.85 | SE 0.08). Das Prinzip der Wortbildung mit insgesamt 71 Struktureinheiten umfasst 32 (45,07 %) leichte und 39 (54,93 %) schwere Einheiten. Betrachtet man die Itemparameter, so zeigt sich, dass 13 Einheiten von den Analysen ausgeschlossen werden müssen. Sowohl das Kriterium für den Item-Fit als auch für die Trennschärfe werden von zwei schweren Einheiten (162, 171) verletzt. Die Trennschärfe unterschreiten neun leichte (127, 131, 132, 133, 138, 142, 145, 180, 184) und zwei schwere Einheiten (151, 170). Die Personenfähigkeit (MW 1.83 | SE 0.07) deutet auf eine gute Schülerkompetenz in diesem Prinzip hin. Die fünfte Teilkompetenz bezogen auf das wortübergreifende Prinzip umfasst 38 Struktureinheiten, wovon 20 (52,63 %) leichte und 18 (47,37 %) schwere Itemschwierigkeiten enthalten sind. Es müssen zwei schwere Einheiten (211, 231) wegen der Verletzung des Kriteriums für den Item-Fit und die Trennschärfe entfernt werden. Weitere fünf leichte (196, 202, 215, 217, 225) und zwei schwere Einheiten (214, 230) weisen eine zu geringe Trenn-

schärfe auf. Die Schülerinnen und Schüler verfügen in diesem Prinzip über eine höhe Personenfähigkeit (MW 2.03 | SE 0.06).

Somit resultiert das Ausgangsmodell mit verbleibenden 200 Struktureinheiten, die sich wie folgt auf die fünf Teilkompetenzen verteilen und Veränderungen in Bezug auf die Personenfähigkeit der Schülerinnen und Schüler mit sich bringen:

Der Kernbereich besteht nun aus 48 Struktureinheiten bei dem phonographisch-silbischen Prinzip und 42 Struktureinheiten bei dem morphologischen Prinzip. Die Personenfähigkeit verringert sich im Kernbereich leicht (MW 2.27/1.67 | SE 0.08/0.07), da die entfernten Struktureinheiten fast ausschließlich leichte Itemschwierigkeiten aufweisen. Die verbleibenden 23 Struktureinheiten im Peripheriebereich bewirken einen Anstieg der Personenfähigkeit (MW 0.99 | SE 0.08), weil die ausgeschlossene Einheit über eine hohe Itemschwierigkeit verfügt. Das Prinzip der Wortbildung umfasst nun 58 Struktureinheiten, die insgesamt eine schwerere Itemschwierigkeit aufweisen, wodurch die Personenfähigkeit (MW 1.67 | SE 0.07) der Schülerinnen und Schüler sinkt. Dagegen verfügen die 29 Struktureinheiten des wortübergreifenden Prinzips über eine geringere Itemschwierigkeit, wodurch die Personenfähigkeit (MW 2.39 | SE 0.08) deutlich ansteigt.

ZUSAMMENFASSUNG DER ERGEBNISSE DER MEHRDIMENSIONALEN SKALIERUNG

Von den 233 Struktureinheiten der fünf Teilkompetenzen, die der sprachsystematische Rechtschreibtest in der ES K5 im Ursprungsmodell enthält, erfüllen 33 Einheiten nicht die statistischen Kriterien für die mehrdimensionale Skalierung.

Tabelle 8.5: Anzahl und statistische Auffälligkeiten der Struktureinheiten – ES K5 – Mehrdimensionale Skalierung

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
	Struktureinheiten				
Ursprungsmodell	51	49	24	71[22]	38
Item-Fit & Trennschärfe	-	-2	-	-2	-2
Item-Fit	-	-	-	-	-
Trennschärfe	-3	-5	-1	-11	-7
Ausgangsmodell	48	42	23	58	29

Konkret müssen 24 (72,73 %) leichte und neun (27,27 %) schwere Struktureinheiten von den Analysen ausgeschlossen werden. Erneut zeigt sich, dass mehr leichte Struktureinheiten von den Analysen ausgeschlossen werden müssen. Dies führt zu dem Ausgangsmodell mit den verbleibenden 200 Struktureinheiten, das bei einem Verlust von 14 Prozent noch eine ausreichende Grundlage für die nachfolgenden Analysen darstellt. Im Gegensatz zur eindimensionalen Skalierung entfallen bei der mehrdimensionalen Skalierung neun weitere Struktureinheiten, weswegen davon ausgegangen werden kann, dass sich mit steigender Komplexität und dem Zusammenspiel der unterschiedlichen Dimensionen der Skalierungsmodelle die Anzahl der auszuschließenden Struktureinheiten erhöht.

Inwiefern sich die Reduzierung der Struktureinheiten auf die Mittelwerte und Standardabweichungen der Personenfähigkeit der jeweiligen Teilkompetenz auswirkt, ist in Tabelle 8.6 zusammengefasst.

[22] Eine Struktureinheit wurde von allen Schülerinnen und Schülern richtig gelöst und wurde direkt ausgeschlossen.

Tabelle 8.6: Veränderungen der deskriptiven Statistiken – ES K5 – Mehrdimensionale Skalierung

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
			Struktureinheiten		
Ursprungsmodell					
MW	2.40	1.80	0.85	1.83	2.03
SE	0.08	0.07	0.08	0.07	0.06
Ausgangsmodell					
MW	2.27	1.67	0.99	1.67	2.39
SE	0.08	0.07	0.08	0.07	0.08
Angabe in Logits für die Personenfähigkeit					

Im Vergleich zu Tabelle 8.3, in der die deskriptiven Werte der eindimensionalen Skalierungen dargestellt sind, fallen hier nur geringfügige Abweichungen auf, die sich erst in der zweiten Nachkommastelle zeigen und der Schätzung bei der Skalierung geschuldet sind. Ansonsten ist auch an dieser Stelle erneut zu erkennen, dass die Teilkompetenzen mit Ausnahme des wortübergreifenden Prinzips im Ausgangsmodell schwerer werden, da die Mittelwerte für die Personenfähigkeiten sinken. Dies trifft nun auch für den Peripheriebereich zu, da bei der eindimensionalen Skalierung keine Struktureinheit entfernt und erst bei der mehrdimensionalen Skalierung eine Struktureinheit auffällig wurde.

Die Reduzierung der Struktureinheiten lässt sich weitergehend hinsichtlich der Werte für die Reliabilitäten und Korrelationen betrachten.

Abbildung 8.10: Latente Korrelationen und Reliabilitäten für das Ursprungs- und Ausgangsmodell – ES K5 – Mehrdimensionale Skalierung

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
Phonographisch-silbischer Kernbereich	0.93 0.94	0.98	0.96	0.97	0.92
Morphologischer Kernbereich	0.98	0.92 0.93	0.99	0.96	0.87
Peripheriebereich	0.96	0.99	0.92 0.92	0.94	0.85
Prinzip der Wortbildung	0.97	0.95	0.95	0.93 0.94	0.94
Wortübergreifendes Prinzip	0.91	0.84	0.83	0.90	0.88 0.86

Latente Interkorrelationen (dargestellt auf der oberen und unteren Dreiecksmatrix) sowie Reliabilitäten (EAP/PV) (dargestellt auf der Diagonalen) der fünf Teilfähigkeiten in der ES K5 für das Ausgangsmodell (obere Angaben) und das Ursprungsmodell (untere Angaben).

Es konnten für das Ursprungs- und Ausgangsmodell Reliabilitäten um 0.90 für die Teilkompetenzen nachgewiesen werden. So liegen die Reliabilitäten für die ersten vier Teilkompetenzen bei dem Ursprungsmodell zwischen 0.92 und 0.94 und verringern sich beim Ausgangsmodell auf 0.92 bzw. 0.93, außer beim Peripherieberiech, der konstant eine Reliabilität in Höhe von 0.92 aufweist. Das wortübergreifende Prinzip hat die verhältnismäßig geringsten Reliabilitäten mit 0.86 bei dem Ursprungsmodell und 0.88 bei dem Ausgangsmodell. Demnach sinken die Reliabitäten geringfügig durch die Reduzierung der Struktureinheiten bei vier von fünf Teilkompetenzen. Aufgrund dieser Ergebnisse ist eine zuverlässige Erfassung der Rechtschreibkompetenz gewährleistet. Auffällig ist, dass die Reliabilitäten im Vergleich zu den eindimensionalen Skalierungen (Kapitel 8.1.1.2) für alle Teilkompetenzen höher sind.

Die Korrelationen sind bei den fünf Teilkompetenzen sowohl für das Ursprungsmodell als auch bei dem Ausgangsmodell auf einem hohen Niveau. Der Kernbereich bestehend aus dem phonographisch-silbischen und morphologischen Prinzip korreliert bei dem Ursprungs- und Ausgangsmodell mit 0.98 und weist damit die zweithöchste Korrelation auf. Der Peripheriebe-

reich korreliert mit 0.99 noch höher mit dem morphologischen Prinzip. Im Fall der drei Teilkompetenzen ist die Eigenständigkeit der Prinzipien als grenzwertig zu bezeichnen. Dagegen korreliert das wortübergreifende Prinzip am geringsten mit den restlichen Teilkompetenzen. Das trifft sowohl für das Ursprungsmodell als auch für das Ausgangsmodell zu. In Hinblick auf die vorhergegangenen Untersuchungen zum sprachsystematischen Rechtschreibtest im Rahmen von IGLU und der HeLp-Studie (Kapitel 5.2) zur Reliabilitäts- und Korrelationsstruktur zeigt sich ein vergleichbares Bild, wobei insbesondere die Korrelationen in Klassenstufe 5 ansteigen und sich immer schlechter ausdifferenzieren lassen. Daher wurden weitere Analysen durchgeführt.

Inwiefern sich das theoretische Kompetenzkonstrukt des sprachsystematischen Rechtschreibtests am besten empirisch modellieren lässt, wird abschließend mit der Vorstellung der Ergebnisse der Modellgeltungstests aufgezeigt.

8.1.1.4 DIMENSIONALITÄT DES KOMPETENZKONSTRUKTS

Tabelle 8.7 sind die vier Modellierungsvarianten der fünf Teilkompetenzen zu entnehmen. Bei dem fünfdimensionalen Modell, bei dem jede Teilkompetenz eine Dimension einnimmt, ist ein Deviance-Wert in Höhe von ca. 46116 ermittelt worden. Die Zusammenführung des Kernbereichs ergibt einen um ca. 15 Punkte höheren Wert. Die weitere didaktisch und statistisch begründbare Reduzierung der Dimensionen auf ein zweidimensionales Modell mit den wortbezogenen Prinzipien (1-4) und dem syntaxbezogenem Prinzip (5) bringt im Vergleich zum fünfdimensionalen Modell einen um 78 Punkte höheren Deviance-Wert.

Tabelle 8.7: Modellgeltungstests – ES K5 – Mehrdimensionale Skalierung

	1D	2D	4D	5D
Deviance	46234.40556	46194.72762	46131.86584	**46116.82433**
AIC	46636.40556	46640.40556	46551.86584	**46546.82433**
CAIC	47580.52135	47593.91554	**47538.25547**	47556.69943
BIC	47379.52135	47390.91554	**47328.25547**	47341.69943

ES K5 (N = 298, 200 Struktureinheiten)

Die Überprüfung der Modellpassung in Bezug auf die Deviance hat ergeben, dass die Repräsentation der Datenstruktur durch das mehrdimensionale Modell mit einem signifikanten Unterschied der Deviance-Werte von 118 Punkten besser als bei einem eindimensionalen Modell gegeben ist. Somit entspricht es dem Ergebnis, das zur Dimensionalität des Kompetenzkonstrukts im Rahmen von IGLU und der HeLp-Studie (Kapitel 5.2) gefunden wurde.

Werden zusätzlich die Paramateranzahl und die Stichprobengröße für die Modellpassung berücksichtigt, zeigt sich bei dem Akaike's Information Criterion (AIC) erneut die bessere Repräsentation der Datenstruktur bei dem fünfdimensionalen Modell. Hingegen liegt nach dem Consistent AIC (CAIC) oder Bayes Information Criterion (BIC) die beste Modellpassung bei dem vierdimensionalen Modell vor.

Demnach kann keine eindeutige Lösung für die beste Modellierung des Kompetenzkonstrukts ermittelt werden. Allerdings deuten alle Maße zur Bewertung der Modellpassung eine mehrdimensionale Datenstruktur an, die im Gegensatz zum eindimensionalen Modell die Informationen in den Daten besser abbildet.

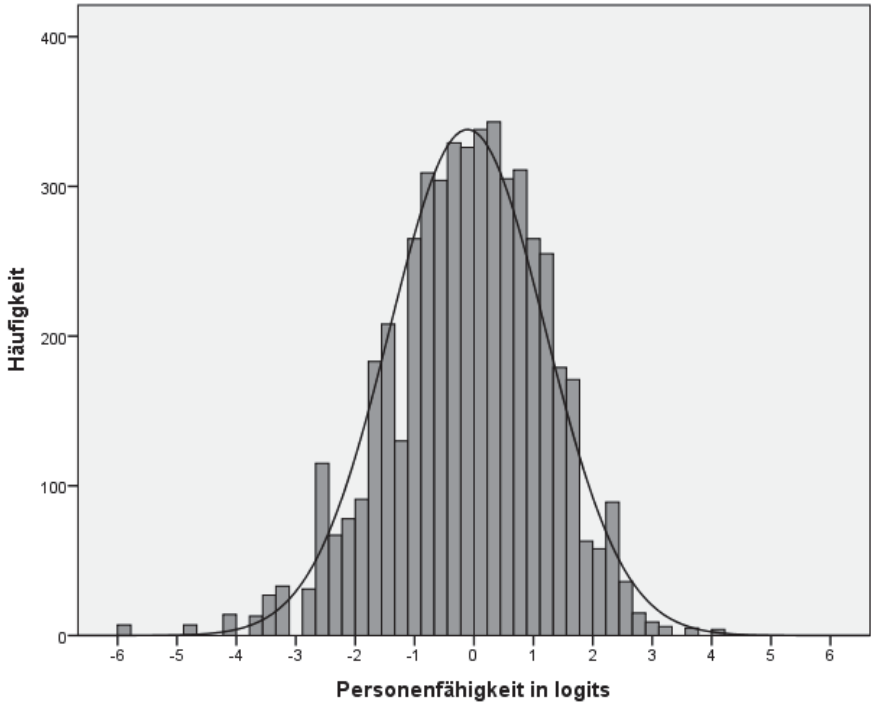
8.1.2 DATENGRUNDLAGE DER HAUPTERHEBUNG K5

Analog zu Kapitel 8.1.1 wird für die Haupterhebung in Klassenstufe 5 (HE K5) einleitend die Datenprüfung veranschaulicht. Nachfolgend werden die Ergebnisse der ein- und mehrdimensionalen Skalierung mittels der Item-, Personen- und Modellparameter vorgestellt, um abschließend die Dimensionalität des Kompetenzkonstrukts zu klären.

8.1.2.1 DATENPRÜFUNG

Zunächst wird auf die Überprüfung einer Normalverteilung für die Item- und Personenparameter im Fall des ganzen Wortes für die HE K5 auf der Grundlage eines Kolmogorov-Smirnov-Tests eingegangen. Abbildung 8.11 stellt in einem Histogramm die geschätzten Personenfähigkeiten in Logits gemeinsam mit der Normalverteilungskurve grafisch dar.

Abbildung 8.11: Verteilung der Personenfähigkeit für das ganze Wort – HE K5 (N = 4.989)



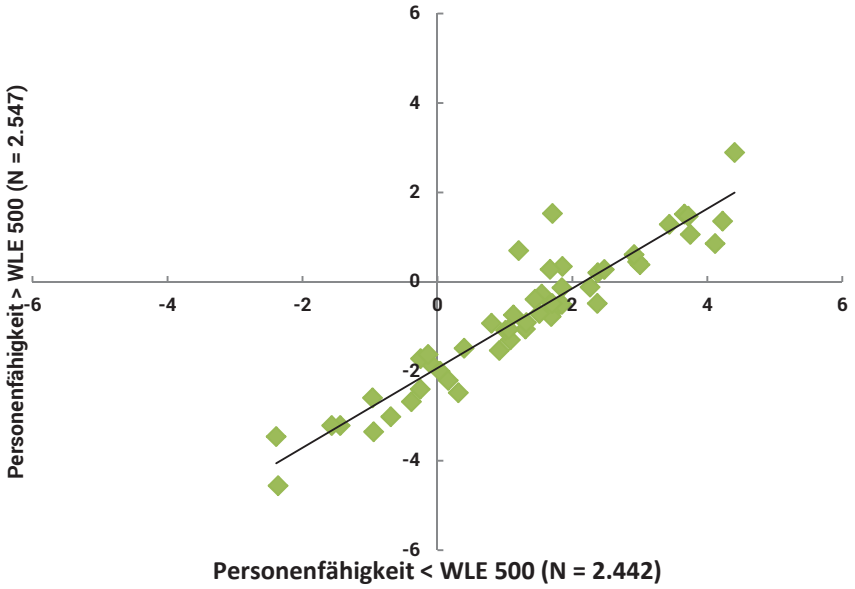
Die Stichprobe der HE K5 weist für die Personenfähigkeit einen Mittelwert von -0.11 Logits und eine Standardabweichung von 1.31 Logits auf. Anhand der statistischen Überprüfung einer

Normalverteilung konnte ein Signifikanzniveau von 0.00 für den Personenparameter ermittelt werden, d. h., dass eine Verzerrung der Personenfähigkeit vorliegt und mehr mittlere und gute Schülerinnen und Schüler an der Testung teilgenommen haben.

Für die Teilkompetenzen wird die Prüfung einer Normalverteilung in der Zusammenfassung (Kapitel 8.1.2.2) behandelt.

Der durchgeführte grafische Modelltest mit den Itemschwierigkeiten der 54 ganzen Wörter ist für die zwei Teilstichproben (Personenfähigkeit < oder > 500) der HE K5 in Abbildung 8.12 dargestellt.

Abbildung 8.12: Grafischer Modelltest auf Basis der Itemschwierigkeiten – HE K5



In Bezug auf die Annahme des Rasch-Modells ist zu erkennen, dass sich Differenzen zwischen 0.18 und 3.26 Logits ergeben. Es wird eine erklärte Varianz (R^2) von 0.97 nachgewiesen.

Durch die Prüfung der Personenparameter auf eine Normalverteilung lässt sich nun festhalten, dass die Datengrundlage in Bezug auf die Personenfähigkeit eine leichte Verzerrung aufweist. Zudem ist wegen des grafischen Modelltests zu folgern, dass sich die Itemschwierigkeiten bei den beiden Teilstichproben unterscheiden. Dies unterstützt den in Kapitel 8.1.1.1 vermerkten Hinweis zur Verletzung der Personenhomogenität und der Annahme einer unterschiedlichen

Personenhomogenität. Aufgrund der rein deskriptiven Beurteilung wird mit der benannten Einschränkung das Rasch-Modell zur Kompetenzmodellierung vorläufig zur weiteren Begutachtung angenommen.

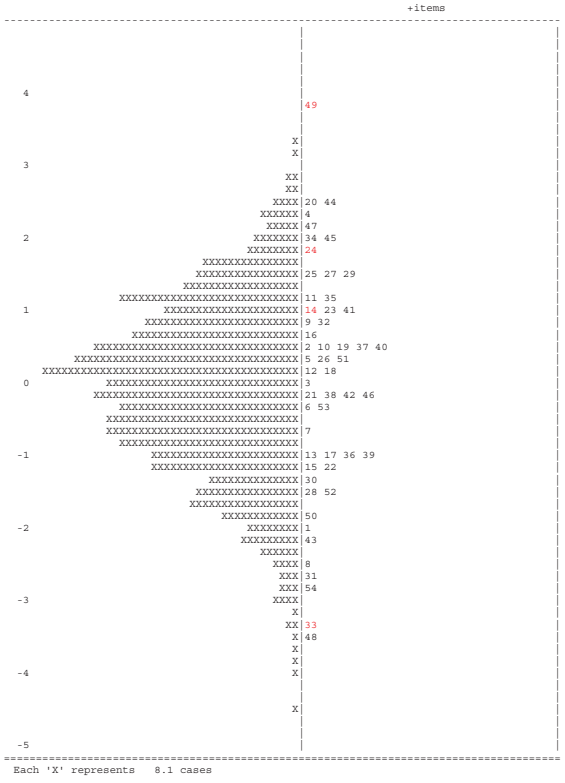
8.1.2.2 EINDIMENSIONALE SKALIERUNGEN

Inwiefern sich bei der HE K5 die Ursprungsmodelle von den optimierten Ausgangsmodellen unterscheiden, wird auf Grundlage der getrennt eindimensional skalierten Daten für das ganze Wort und die Struktureinheiten veranschaulicht. Die dabei geschätzten Itemschwierigkeiten und Personenfähigkeiten sind in den Abbildung 8.13 bis Abbildung 8.18 auf einer gemeinsamen Skala abgetragen.

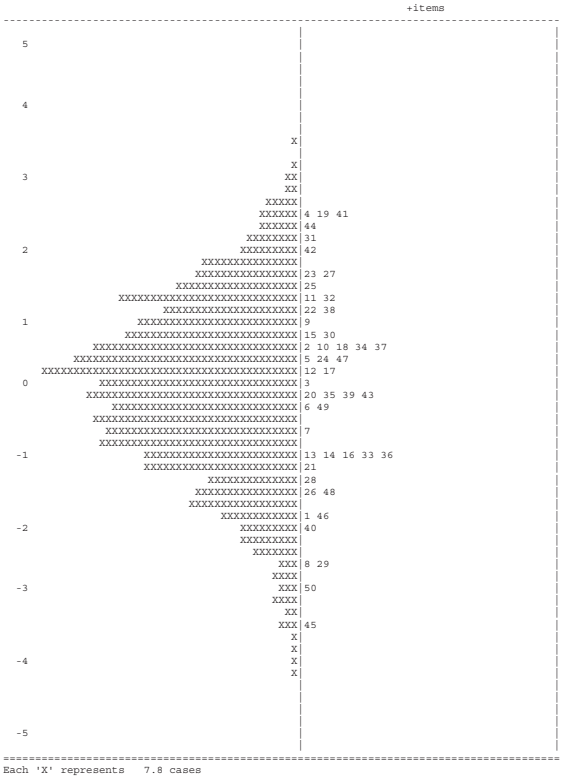
Einleitend erfolgt die Gegenüberstellung der beiden Skalierungsmodelle für das ganze Wort der HE K5.

Abbildung 8.13: Ganzes Wort – HE K5

Ursprungsmodell



Ausgangsmodell

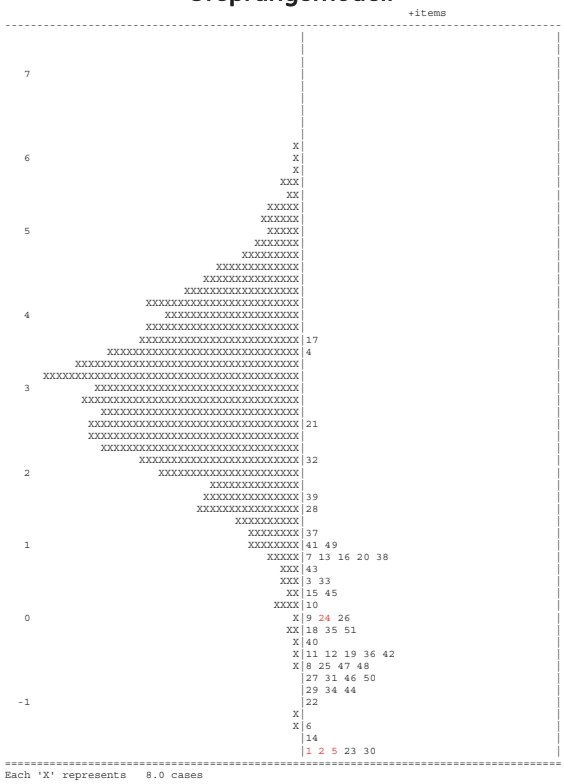


Bei den 54 ganzen Wörtern lassen sich 25 (46,30 %) leichte bzw. 29 (53,70 %) schwere Wörter identifizieren. Statistisch auffällig sind insgesamt fünf Wörter, wobei zwei schwere Wörter (14, 24) sowohl das Kriterium für den Item-Fit (>1.20) als auch das der Trennschärfe (<0.26) verletzen. Jeweils ein leichtes und schweres Wort (33, 49) weisen eine zu geringe Trennschärfe auf. Die Personenfähigkeit (MW -0.11 | SE 0.02) deutet auf eine verringerte Schülerfähigkeit hin. Das Ausgangsmodell mit den verbleibenden 50 Wörtern weist eine leicht gestiegene Personenfähigkeit (MW -0.05 | SE 0.02) auf.

In einer identischen Darstellungsform folgen nun die Ergebnisse für den Kernbereich bestehend aus dem phonographisch-silbischen sowie dem morphologischen Prinzip.

Abbildung 8.14: Phonographisch-silbisches Prinzip im Kernbereich – HE K5

Ursprungsmodell



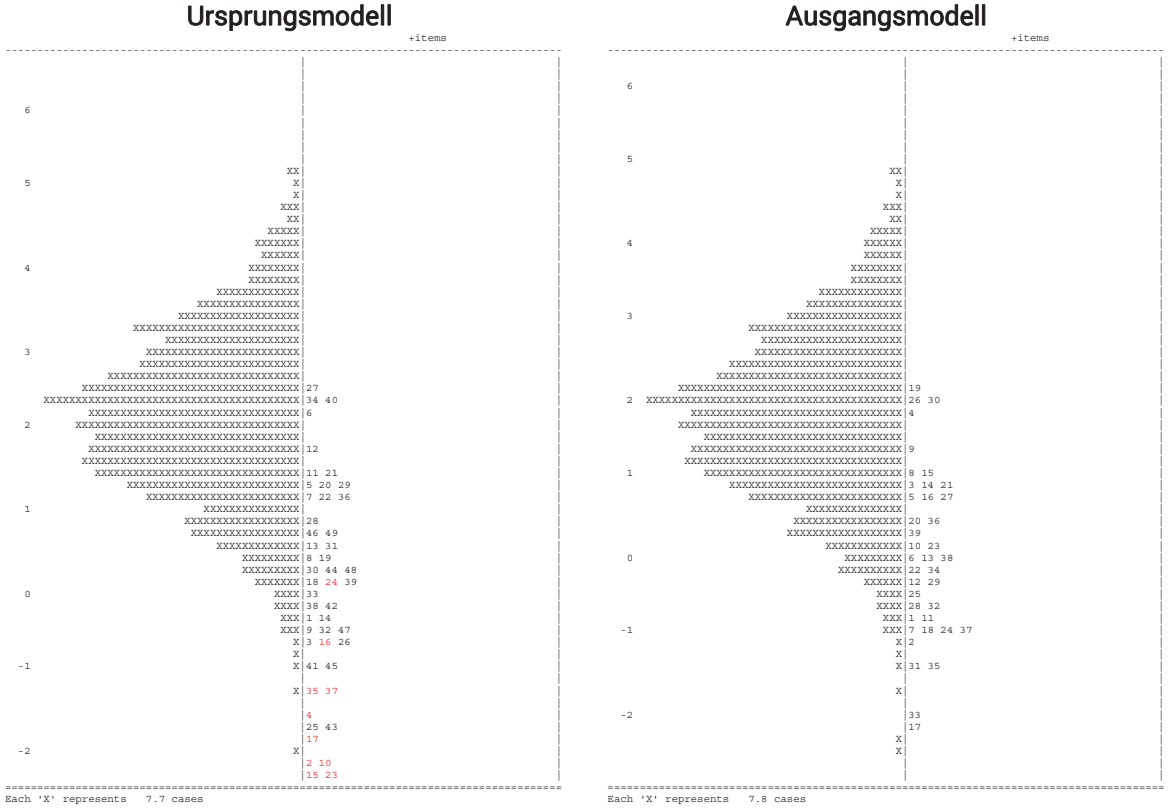
Ausgangsmodell

Das Ursprungsmodell des phonographisch-silbischen Prinzips enthält 51 Struktureinheiten, die sich in 29 (56,86 %) leichte bzw. 22 (43,14 %) schwere Einheiten differenzieren lassen. Es müssen drei leichte Einheiten (1, 2, 5) und eine schwere Einheit (24) wegen einer zu geringen

Trennschärfe von den Analysen ausgeschlossen werden. Auf höhere Schülerfähigkeiten in dieser Teilkompetenz deutet die Personenfähigkeit (MW 2.90 | SE 0.02) hin. Es ergibt sich das Ausgangsmodell mit insgesamt 47 Struktureinheiten, wobei es zu einer Verringerung der Personenfähigkeit (MW 2.79 | SE 0.02) kommt.

Weitergehend werden Ergebnisse der deskriptiven Analyse der Item- und Personenparameter zum morphologischen Prinzip des Kernbereichs referiert.

Abbildung 8.15: Morphologisches Prinzip im Kernbereich – HE K5

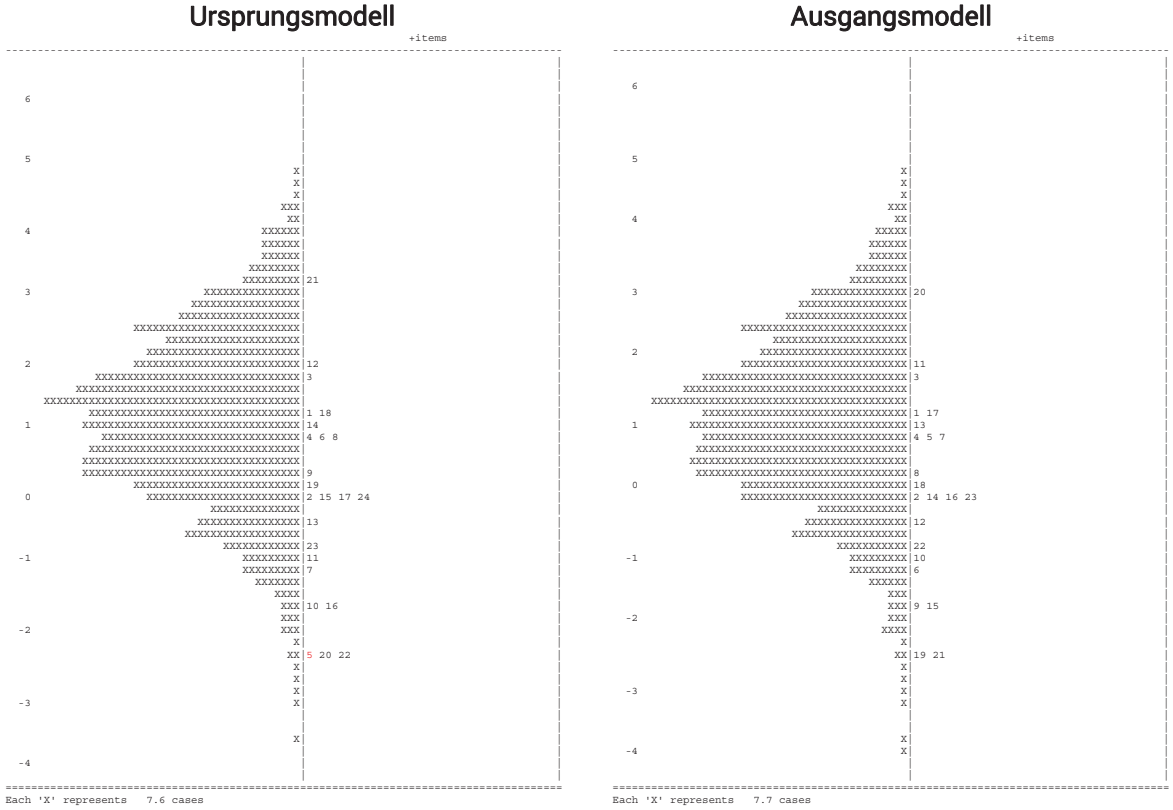


Das morphologische Prinzip umfasst im Ursprungsmodell 49 Struktureinheiten, wovon 23 (46,94 %) leichte und 26 (53,06 %) schwere Einheiten vorliegen. Es verletzten zehn leichte Einheiten (2, 4, 10, 15, 16, 17, 23, 24, 35, 37) das Kriterium für die Trennschärfe. Die Personenfähigkeit (MW 2.11 | SE 0.02) deutet daraufhin, dass die Schülerinnen und Schüler in der morphologischen Teilkompetenz kompetent sind. Durch das Ausgangsmodell mit den verbleiben-

den 39 Struktureinheiten sinkt die Personenfähigkeit (MW 1.71 | SE 0.02) der Schülerinnen und Schüler.

Analog erfolgt nun die deskriptive Darstellung des Peripheriebereichs, wie es in Abbildung 8.16 veranschaulicht ist.

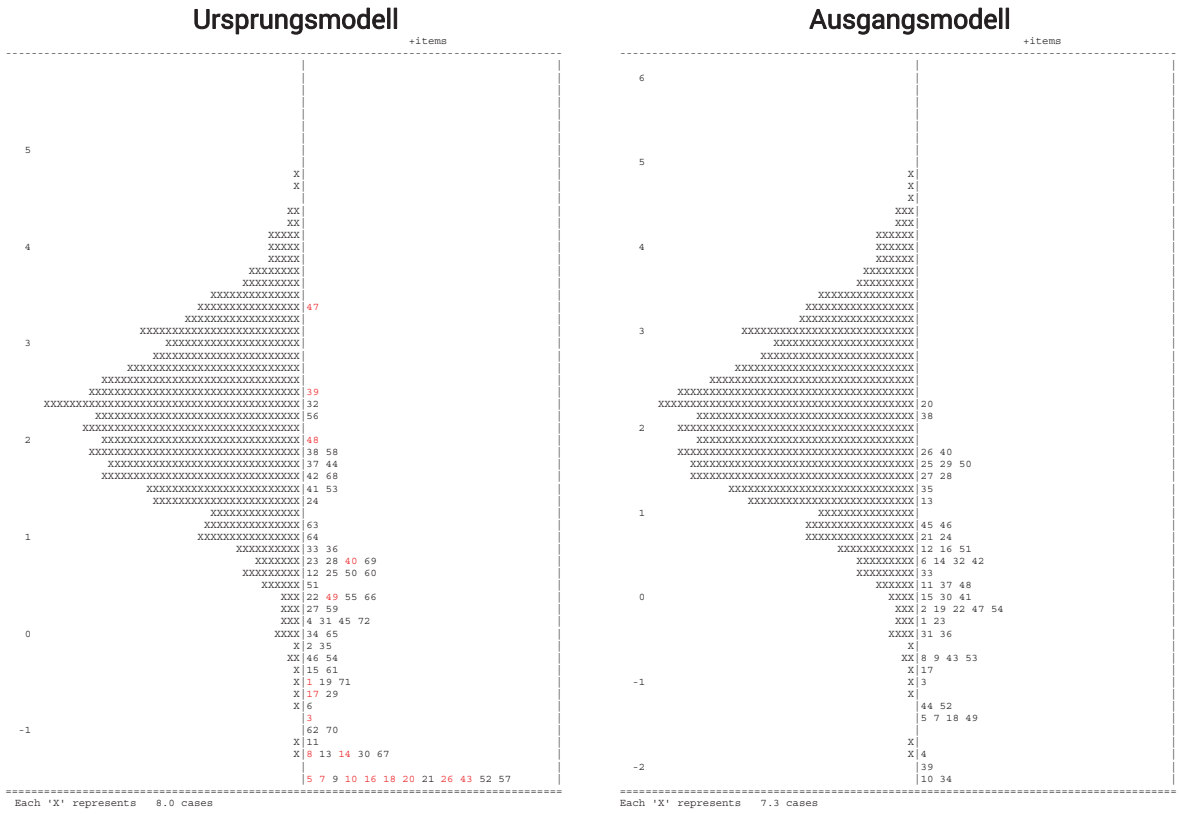
Abbildung 8.16: Peripheriebereich – HE K5



Das Ursprungsmodell des Peripheriebereichs umfasst 24 Struktureinheiten, wovon sich zehn (41,67 %) leichte und 14 (58,33 %) schwere Einheiten identifizieren lassen. Lediglich eine leichte Einheit (5) unterschreitet das statistische Kriterium für die Trennschärfe und wird von den Analysen ausgeschlossen. Die Personenfähigkeit (MW 1.14 | SE 0.02) weist im Vergleich zu den anderen Teilkompetenzen die geringste Schülerfähigkeit auf. Es verbleiben 23 Struktureinheiten bei dem Ausgangsmodell, wobei eine leicht gesunkene Personenfähigkeit (MW 1.03 | SE 0.02) zu verzeichnen ist.

Nachfolgend wird das Prinzip der Wortbildung anhand der deskriptiven Analyse für die Item- und Personenparameter dargestellt.

Abbildung 8.17: Prinzip der Wortbildung – HE K5

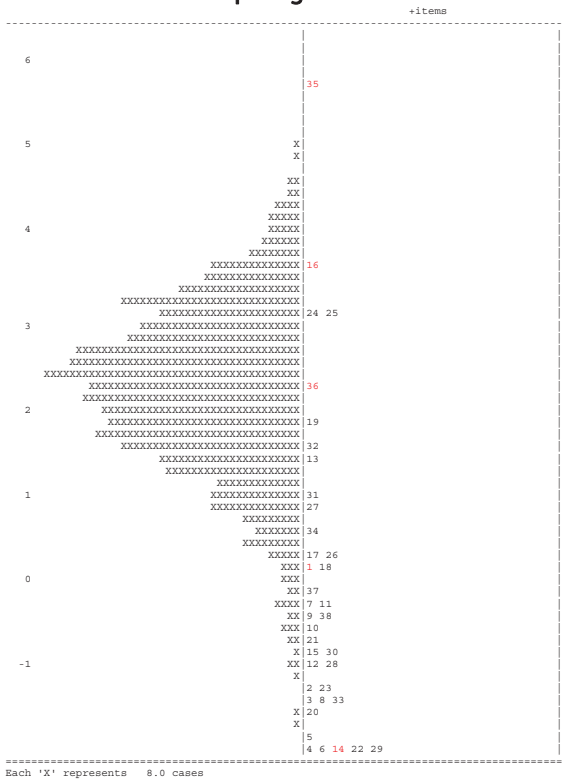


Bei dem Prinzip der Wortbildung liegen 72 Struktureinheiten im Ursprungsmodell vor, die 33 (45,83 %) leichte und 39 (54,17 %) schwere Itemschwierigkeiten aufweisen. Davon sind 18 Struktureinheiten, und zwar zwölf leichte und sechs schwere Einheiten, statistisch auffällig. Eine schwere Einheit (39) verletzt sowohl das Kriterium für den Item-Fit als auch das Kriterium für die Trennschärfe. Wiederum eine schwere Einheit (40) überschreitet die Grenze für den Item-Fit. Die restlichen zwölf leichten Einheiten (1, 3, 7, 8, 10, 14, 16, 17, 18, 20, 26, 43) und vier schwere Einheiten (5, 47, 48, 49) weisen eine zu geringe Trennschärfe auf. Die Personenfähigkeit (MW 2.19 | SE 0.01) deutet auf eine hohe Schülerkompetenz in diesem Prinzip hin. Das Ausgangsmodell mit 54 Struktureinheiten führt zu einer leicht gesunkenen Personenfähigkeit (MW 2.00 | SE 0.02).

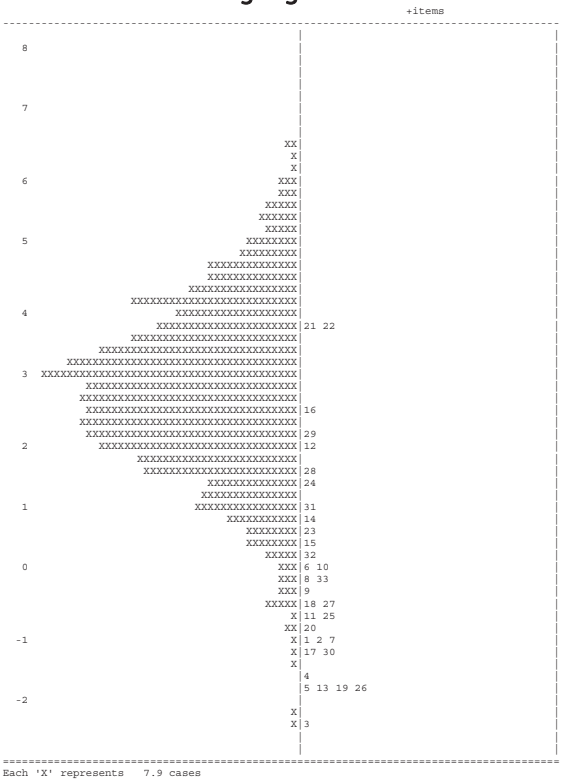
Abschließend erfolgt die deskriptive Darstellung des wortübergreifenden Prinzips in der bekannten Vorgehensweise.

Abbildung 8.18: Wortübergreifendes Prinzip – HE K5

Ursprungsmodell



Ausgangsmodell



Das Ursprungsmodell des wortübergreifenden Prinzips umfasst 38 Struktureinheiten, wovon sich 24 (63,16 %) leichte und 14 (36,84 %) schwere Wörter identifizieren lassen. Fünf Struktureinheiten verletzen die statistischen Kriterien und werden von den Analysen ausgeschlossen. Konkret wird eine leichte Einheit (36) ausgeschlossen, weil sie die Grenzen für den Item-Fit und der Trennschärfe überschreitet. Eine weitere schwere Einheit (1) weist einen zu hohen Item-Fit auf. Die restlichen drei auffälligen Struktureinheiten, eine leichte Einheit (15) und zwei schwere Einheiten (16, 35), besitzen eine zu geringe Trennschärfe. Die Personenfähigkeit (MW 2.17 | SE 0.02) der Schülerinnen und Schüler zeigt eine hohe Kompetenz in diesem Prinzip. Bei dem resultierenden Ausgangsmodell verbleiben 33 Struktureinheiten. Die Personenfähigkeit (MW 2.70 | SE 0.02) steigt an, da überwiegend schwere Einheiten entfernt werden.

Die Erfassung des differenziellen Rechtschreibkompetenzmodells in Form der Reliabilitäten ist in Tabelle 8.8 zusammengetragen.

Tabelle 8.8: Reliabilitäten – HE K5 – Eindimensionale Skalierungen

	Ganzes Wort	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
	Wörter	Struktureinheiten				
Ursprungsmodell	0.92	0.79	0.84	0.85	0.84	0.78
Ausgangsmodell	0.92	0.79	0.84	0.85	0.84	0.80

Sowohl die Ursprungs- als auch die Ausgangsmodelle gewährleisten eine zuverlässige Erfassung der Rechtschreibung, da die ermittelten Reliabilitäten alle über 0.70 liegen. Bei den ganzen Wörtern zeigt sich im Gegensatz zu den Teilkompetenzen, dass hier höhere Reliabilitäten festgestellt werden. Die Reduzierung der statistisch auffälligen Wörter bzw. Struktureinheiten bewirkt nur beim wortübergreifenden Prinzip eine geringe Veränderung der Reliabilität um 0.02, da sie ansonsten unverändert bleibt. Der sprachsystematische Rechtschreibtest ist somit im Rahmen der HE K5 in der Lage, das differenzielle Rechtschreibkompetenzmodell in der Klassenstufe 5 auf eindimensionaler Ebene zu erfassen.

ZUSAMMENFASSUNG DER EINDIMENSIONALEN SKALIERUNGEN

Für die HE K5 lässt sich in Tabelle 8.9 zusammenfassen, dass vier von den ursprünglich 54 ganzen Wörtern und 38 von den 234 Struktureinheiten aus statistischen Gründen nicht für die kompetenzorientierte Leistungsmessung geeignet sind.

Tabelle 8.9: Anzahl und statistische Auffälligkeiten der ganzen Wörter und Struktureinheiten – HE K5 – Eindimensionale Skalierungen

	Ganzes Wort	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
	Wörter	Struktureinheiten				
Ursprungsmodell	54	51	49	24	72	38
Item-Fit & Trennschärfe	-2	-	-	-	-1	-1
Item-Fit	-	-	-	-	-1	-1
Trennschärfe	-2	-4	-10	-1	-16	-3
Ausgangsmodell	50	47	39	23	54	33

Die verbleibenden 50 ganzen Wörter und 196 Struktureinheiten für die Datenanalyse der eindimensionalen Modelle bedeuten einen Verlust von 7.41 bzw. 16.24 Prozent. Folglich stehen für die Datenanalyse und Weiterentwicklung des Testinstruments mehr als 83 Prozent der eingesetzten Auswertungseinheiten zur Verfügung, d. h., es besteht eine ausreichende Auswertungsgrundlage.

Die deskriptiven Statistiken in Tabelle 8.10 zeigen für die Ursprungsmodelle auf der Ebene des ganzen Wortes, dass das eingesetzte Testinstrument im Verhältnis zur ermittelten Personenfähigkeit etwas zu schwer ist und die Kompetenz minimal verringert ist. Die Ebene der Struktureinheiten der Teilkompetenzen weist mit Ausnahme des Peripheriebereichs eine hohe Kompetenz der Schülerinnen und Schüler auf. Die Normalverteilung der Personenparameter kann bei keinem der Ursprungsmodelle nachgewiesen werden.

Tabelle 8.10: Veränderungen der deskriptiven Statistiken – HE K5 – Eindimensionale Skalierungen

	Ganzes Wort	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
	Wörter	Struktureinheiten				
Ursprungsmodell						
MW	-0.11/0.00	2.90/0.00	2.11/0.00	1.14/0.00	2.19/0.00	2.17/0.00
SE	0.02/0.22	0.02/0.18	0.02/0.19	0.02/0.30	0.01/0.17	0.02/0.29
NV	×/-	×/-	×/-	×/-	×/-	×/-
Ausgangsmodell						
MW	-0.05/0.00	2.79/0.00	1.71/0.00	1.03/0.00	2.00/0.00	2.70/0.00
SE	0.02/0.21	0.02/0.19	0.02/0.17	0.02/0.29	0.02/0.15	0.02/0.27
NV	×/-	×/-	×/-	×/-	×/-	×/-
Angabe in Logits für die Personenfähigkeit/Itemschwierigkeit						

Die Entfernung der statistisch auffälligen ganzen Wörter bzw. Struktureinheiten bewirkt bei den Ausgangsmodellen größtenteils eine verringerte Personenfähigkeit, da überwiegend leichte Wörter bzw. Einheiten von den Analysen ausgeschlossen werden. Dies trifft nicht für das ganze Wort und das wortübergreifende Prinzip zu, wo es zu einer Steigerung der Personenfähigkeit kommt. Die Standardfehler der Personenfähigkeit sind konstant, wobei sie bei der Itemschwierigkeit mit Ausnahme beim phonographisch-silbischen Prinzip sinken. Die Normalverteilung liegt weiterhin nicht vor. Fest steht, dass die Schülerinnen und Schüler häufiger einzelne Struktureinheiten einer Teilkompetenz als ein ganzes Wort korrekt schreiben.

Abschließend wird in Tabelle 8.11 dargestellt, inwiefern es durch die Optimierung der Ursprungsmodelle Veränderungen der prozentualen Lösungshäufigkeit bei den ganzen Wörtern und den Struktureinheiten gibt.

Tabelle 8.11: Prozentualen Lösungshäufigkeiten – HE K5

		Ganzes Wort	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
		Wörter	Struktureinheiten				
Ursprungsmodell	MW	48.21	88.24	80.69	67.02	81.94	80.03
	(SE)	(0.27)	(0.15)	(0.20)	(0.29)	(0.16)	(0.17)
Ausgangsmodell	MW	49.15	87.52	77.04	65.83	80.83	84.51
	(SE)	(0.29)	(0.16)	(0.23)	(0.30)	(0.20)	(0.19)

Angaben in Prozent

Beim Vergleich der Ursprungsmodelle ist zwischen dem ganzen Wort und den Teilkompetenzen zu erkennen, dass die prozentuale Lösungshäufigkeit der Schülerinnen und Schüler auf der Ebene der Teilkompetenzen deutlich höher ist. Während die Lösungshäufigkeit bei den ganzen Wörtern knapp unterhalb von 50 Prozent liegt, bewegt sie sich bei den Teilkompetenzen zwischen 65 und 88 Prozent. Am besten beherrschen die Schülerinnen und Schüler das phonographisch-silbische Prinzip, wo die Lösungshäufigkeit bei ungefähr 88 Prozent liegt. Der Peripheriebereich weist die geringste Lösungshäufigkeit auf, wobei die übrigen Teilkompetenzen zu knapp 81 Prozent richtig geschrieben werden. Durch die Reduzierung der statistisch auffälligen Wörter bzw. Struktureinheiten sinkt bei den Ausgangsmodellen größtenteils die prozentuale Lösungshäufigkeit, wobei sie beim ganzen Wort und beim wortübergreifenden Prinzip steigt.

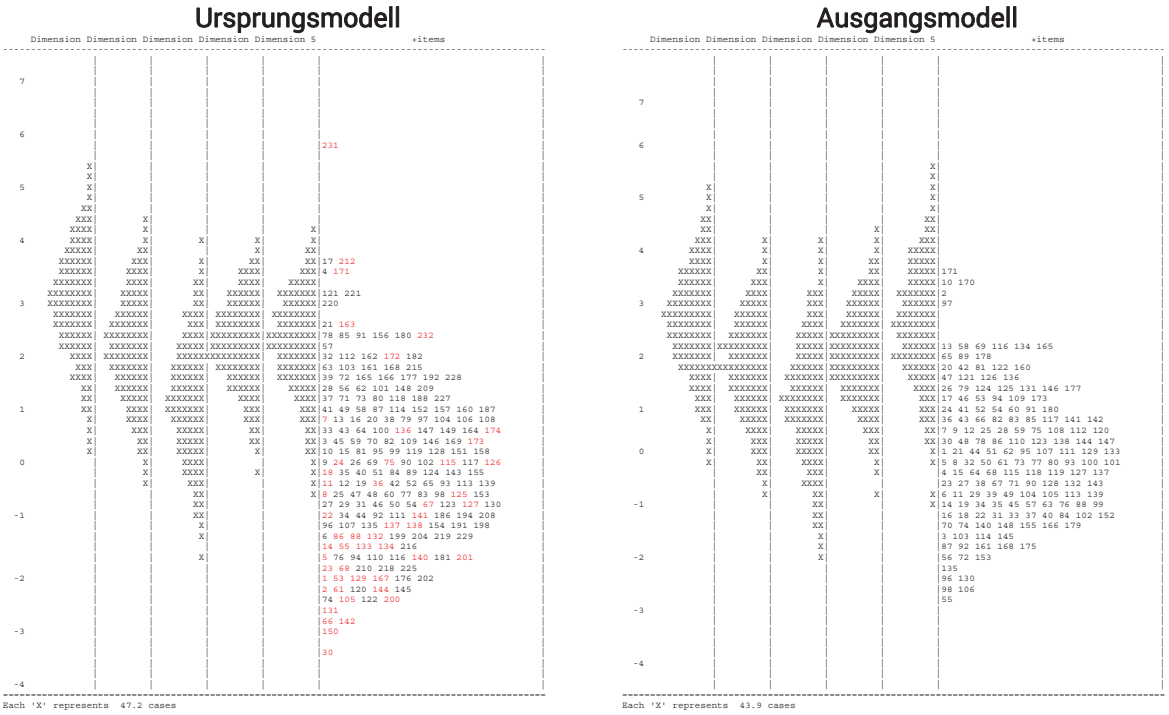
Nachfolgend werden die Daten aus der HE K5 mehrdimensional betrachtet und die Ergebnisse zu den Reliabilitäten und Korrelationen auf der Ebene der Teilkompetenzen berichtet.

8.1.2.3 MEHRDIMENSIONALE SKALIERUNG

Analog zur eindimensionalen Betrachtung wird einleitend eine umfassende Analyse der Item-Personen- und Modellparameter auf Grundlage der mehrdimensionalen Skalierung vorgestellt.

Anhand der Gegenüberstellung des Ursprungs- und Ausgangsmodells der fünfdimensionalen Skalierung der fünf Teilkompetenzen werden nachfolgend die Ergebnisse zu den Item- und Personenparametern behandelt.

Abbildung 8.19: Mehrdimensionale Skalierung – HE K5



234 Struktureinheiten umfasst das Ursprungsmodell der fünfdimensionalen Skalierung, die sich wie folgt auf die fünf Teilkompetenzen verteilen:

In den ersten beiden Spalten bzw. Dimensionen auf der linken Seite der Abbildung 8.19 sind der Kernbereich, bestehend aus dem phonographisch-silbischen Prinzip und dem morphologischen Prinzip, abgebildet. Das phonographisch-silbische Prinzip besteht aus 51 Struktureinheiten, wovon sich 29 (56,86 %) leichte und 22 (43,14 %) schwere Einheiten identifizieren lassen. Das morphologische Prinzip umfasst 49 Struktureinheiten, die sich in 23 (46,94 %) leichte und 26 (53,06 %) schwere Einheiten aufteilen. Im Kernbereich sind insgesamt 22 Struktureinheiten statistisch auffällig. Im Fall des phonographisch-silbischen Prinzips wird das Kriterium für die Trennschärfe von elf leichten Einheiten (1, 2, 5, 8, 11, 14, 18, 22, 23, 30, 36) und zwei schweren Einheiten (7, 24) unterschritten. Beim morphologischen Prinzip wird das Kriterium für die Trennschärfe von acht leichten Einheiten (53, 55, 61, 66, 67, 68, 86,88) und einer schweren Einheit (75) unterschritten. Die Personenfähigkeit (MW 2.89/2.10 | SE 0.02/0.02) weist bei beiden Prinzipien eine hohe Fähigkeit der Schülerinnen und Schüler nach. Mittig befindet sich die Personenfähigkeit des Peripheriebereichs mit 24 Struktureinheiten, die elf (45,83 %) leichte und 13 (54,17 %) schwere Itemschwierigkeiten enthalten. Lediglich eine leichte Einheit (105)

weist eine zu geringe Trennschärfe und eine weitere schwere Einheit (115) weist einen zu hohen Item-Fit auf, weshalb sie von den Analysen ausgeschlossen werden. Die Personenfähigkeit (MW 1.13 | SE 0.02) bestätigt bei diesem Prinzip die geringste Kompetenz. Das Prinzip der Wortbildung umfasst 72 Struktureinheiten und beinhaltet 33 (45,83 %) leichte und 39 (54,17 %) schwere Einheiten. Hinsichtlich des Item-Fits und der Trennschärfe verletzen 22 Struktureinheiten die festgelegten Grenzen. Sowohl der Item-Fit als auch die Trennschärfe sind bei zwei schweren Einheiten (163, 172) auffällig. Die Trennschärfe unterschreiten 16 leichte Einheiten (125, 126, 127, 129, 131, 132, 133, 134, 137, 138, 140, 141, 142, 144, 150, 167) und vier schwere Einheiten (136, 171, 173, 174). Eine hohe Kompetenz in diesem Prinzip zeigt die Personenfähigkeit (MW 2.18 | SE 0.01) der Schülerinnen und Schüler. Das auf der rechten Seite dargestellte wortübergreifende Prinzip besteht aus insgesamt 38 Struktureinheiten, die sich in 23 (60,53 %) leichte und 15 (39,47 %) schwere Einheiten differenzieren lassen. Es müssen fünf Einheiten aufgrund der Verletzung der statistischen Kriterien ausgeschlossen werden. Das Kriterium für den Item-Fit und die Trennschärfe wird von einer schweren Einheit (232) verletzt. Eine schwere Einheit (212) überschreitet die Grenze für den Item-Fit. Drei leichte Einheiten (198, 200, 201) und eine schwere Einheit (231) weisen eine zu geringe Trennschärfe auf. In diesem Prinzip liegt erneut eine hohe Personenfähigkeit vor (MW 2.15 | SE 0.02).

Dies führt zum Ausgangsmodell der mehrdimensionalen Skalierung mit den verbleibenden 182 Struktureinheiten:

Auf der Grundlage der restlichen 38 Struktureinheiten zum phonographisch-silbischen Prinzip resultiert eine gesunkene Personenfähigkeit (MW 2.57 | SE 0.02), da die Mehrzahl der entfernten Einheiten eine leichte Itemschwierigkeit besitzt. Das morphologische Prinzip umfasst noch 40 Struktureinheiten, wobei ebenfalls die Personenfähigkeit (MW 1.78 | SE 0.02) sinkt. Etwas geringer ist die Personenfähigkeit (MW 1.04 | SE 0.02) im Peripheriebereich mit 22 Struktureinheiten. Das Prinzip der Wortbildung mit 50 verbleibenden Einheiten weist eine gesunkene Personenfähigkeit (MW 1.94 | SE 0.02) auf. Bei dem wortübergreifenden Prinzip mit den verbleibenden 32 Struktureinheiten steigt dagegen die mittlere Personenfähigkeit (MW 2.53 | SE 0.02).

ZUSAMMENFASSUNG DER MEHRDIMENSIONALEN SKALIERUNG

52 der 233 Struktureinheiten, die das Rechtschreibkompetenzmodell für die fünf Teilkompetenzen in der HE K5 im Ursprungsmodell enthält, erfüllen nicht die statistischen Kriterien für die mehrdimensionale Skalierung.

Tabelle 8.12: Anzahl und statistische Auffälligkeiten der Struktureinheiten – HE K5 – Mehrdimensionale Skalierung

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
			Struktureinheiten		
Ursprungsmodell	51	49	24	72	38
Item-Fit & Trennschärfe	-13	-	-	-2	-1
Item-Fit	-	-	-1	-	-1
Trennschärfe	-	-9	-1	-20	-4
Ausgangsmodell	38	40	22	50	32

Das Ausgangsmodell mit den verbleibenden 182 Struktureinheiten unterliegt einem Verlust an Einheiten in Höhe von 22,22 Prozent. Somit bleibt eine ausreichende Analysegrundlage von ungefähr 78 Prozent bestehen. Dieser Anteil ist im Gegensatz zur eindimensionalen Skalierung der HE K5 nochmals geringer, weswegen die Vermutung bezüglich der Komplexität der Skalierungsmodelle und der Höhe der auszuschließenden Struktureinheiten erneut zutrifft und zudem von der Stichprobengröße abhängig sein muss. Mit steigender Fallzahl zeigen sich mehr statistische Auffälligkeiten bei der mehrdimensionalen Skalierung.

Die Veränderung vom Ursprungs- zum Ausgangsmodell ist in Tabelle 8.13 hinsichtlich der Mittelwerte und Standardfehler der Personenfähigkeit pro Teilkompetenz nochmals zusammengefasst.

Tabelle 8.13: Veränderungen der deskriptiven Statistiken – HE K5 – Mehrdimensionale Skalierung

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
			Struktureinheiten		
Ursprungsmodell					
MW	2.89	2.10	1.13	2.18	2.15
SE	0.02	0.02	0.02	0.01	0.02
Ausgangsmodell					
MW	2.57	1.78	1.04	1.94	2.53
SE	0.02	0.02	0.02	0.02	0.02
Angabe in Logits für die Personenfähigkeit					

Zwischen der eindimensionalen und mehrdimensionalen Skalierung (vgl. Kapitel 8.1.2.2) bestehen nur geringe Unterschiede im Hinblick auf die ermittelten Mittelwerte und Standardabweichungen für die jeweilige Personenfähigkeit in den fünf Teilkompetenzen. Aufgrund der aus statistischen Gründen bedingten Reduzierung der Struktureinheiten bei den Ursprungsmodellen kommt es bei den restlichen Teilkompetenzen mit Ausnahme des wortübergreifenden Prinzips zu einer Verringerung der mittleren Personenfähigkeit, d. h., dass die jeweilige Kompetenz schwieriger wird.

Inwiefern sich die Veränderungen auch auf die Reliabilitäten und Korrelationen der einzelnen Kompetenzdimensionen auswirken, wird nun vorgestellt.

Abbildung 8.20: Latente Korrelationen und Reliabilitäten der fünf Teilkompetenzen für das Ursprungs- und Ausgangsmodell – HE K5

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
Phonographisch-silbischer Kernbereich	0.93 0.92	0.97	0.95	0.97	0.82
Morphologischer Kernbereich	0.97	0.94 0.94	0.99	0.97	0.79
Peripheriebereich	0.96	0.99	0.93 0.93	0.96	0.79
Prinzip der Wortbildung	0.96	0.95	0.95	0.93 0.92	0.83
Wortübergreifendes Prinzip	0.81	0.78	0.79	0.83	0.85 0.82

Latente Interkorrelationen (dargestellt auf der oberen und unteren Dreiecksmatrix) sowie Reliabilitäten (EAP/PV) (dargestellt auf der Diagonalen) der fünf Teilfähigkeiten in der HE K5 für das Ausgangsmodell (obere Angaben) und das Ursprungsmodell (untere Angaben).

Bei dem Ursprungsmodell konnte die geringste und höchste Reliabilität in Höhe von 0.82 bzw. 0.94 für die fünf Teilkompetenzen ermittelt werden. Es ergeben sich bei den Reliabilitäten des Ausgangsmodells maximale Änderungen in Höhe von 0.03, die sich letztlich zwischen 0.85 und 0.94 bewegen. Insgesamt liegen alle Reliabilitäten der mehrdimensionalen Skalierung auf einem hohen Niveau und gewährleisten eine zuverlässige Erfassung der Rechtschreibkompetenz, die im Vergleich zur eindimensionalen Skalierung (Kapitel 8.1.2.2) noch höher sind.

Es bestehen sowohl für das Ursprungs- als auch für das Ausgangsmodell sehr hohe korrelative Zusammenhänge der fünf Teilkompetenzen, die Korrelationskoeffizienten zwischen 0.78 und 0.99 aufweisen. Deutlich grenzt sich hierbei das wortübergreifende Prinzip von den restlichen Teilkompetenzen ab, da hier der Zusammenhang lediglich auf einem geringen Niveau zwischen 0.78 und 0.83 schwankt. Die übrigen Teilkompetenzen korrelieren sehr stark miteinander, weshalb die Eigenständigkeit dieser Prinzipien fraglich ist. Mit Bezug auf die dargestellten Ergebnisse zur Reliabilitäts- und Korrelationsstruktur zeigt sich im Fall der HE K5 ein vergleichbares Muster.

8.1.2.4 DIMENSIONALITÄT DES KOMPETENZKONSTRUKTS

Die empirische Modellierbarkeit des theoretischen Kompetenzkonstrukts des sprachsystematischen Rechtschreibtests wird am Ende des Kapitels mithilfe von Modellgeltungstests auf Basis der Deviance-Statistik, Akaike's Information Criterion (AIC), Consistent AIC (CAIC) sowie Bayes Information Criterion (BIC) vorgestellt.

Tabelle 8.14 sind vier Varianten zu entnehmen, um die fünf Teilkompetenzen des Kompetenzkonstrukts empirisch zu modellieren. Auf der Ebene der Deviance-Statistik liegt bei dem fünfdimensionalen Modell die beste Modellpassung vor.

Tabelle 8.14: Modellgeltungstests – HE K5 – Mehrdimensionale Skalierung

	1D	2D	4D	5D
Deviance	665319.74747	662704.11856	661618.79388	**661366.40395**
AIC	665685.74747	663074.11856	662002.79388	**661760.40395**
CAIC	667060.99078	664464.39185	663445.67211	**663240.85713**
BIC	666877.99078	664279.39185	663253.67211	**663043.85713**

HE K5 (N = 4.989, 181 Struktureinheiten)

Dies zeigt sich ebenfalls bei den drei Tests AIC, CAIC sowie BIC, die zusätzlich die Parameteranzahl und Stichprobengröße berücksichtigen. Anhand der Ergebnisse zur Testung der Modellpassung anhand der Daten der HE K5 können die Befunde der IGLU und der HeLp-

Studie repliziert werden.[23] Demnach liegt die beste Repräsentation der Datenstruktur durch das mehrdimensionale Modell vor.

8.1.3 VERGLEICH DER ENTWICKLUNGSSTUDIE K5 MIT DER HAUPTERHEBUNG K5

Die Entwicklungsstudie (ES) und Haupterhebung (HE) in Klassenstufe 5 (K5) werden nun hinsichtlich der in den Kapiteln 8.1.1 und 8.1.2 dargestellten Ergebnisse verglichen. Dazu wird der Vergleich in vier Schritten aufgezeigt. In einem ersten Schritt werden die Ergebnisse der Datenprüfungen für die ES und HE K5 gegenübergestellt. Daran knüpft ein umfassender Vergleich der ein- und mehrdimensionalen Skalierungen auf der Grundlage der jeweiligen Ursprungsmodelle an, da nur sie eine vergleichbare Grundlage wegen der identischen Anzahl an ganzen Wörtern und Struktureinheiten bieten. Den letzten Schritt stellt die Gegenüberstellung der Ergebnisse zur Dimensionalität des Kompetenzkonstrukts dar.

8.1.3.1 DATENPRÜFUNG

Die Daten der Schülerinnen und Schüler aus der Entwicklungsstudie (ES) und der Haupterhebung (HE) in Klassenstufe 5 (K5) zeigen in Bezug auf die Normalverteilung der Personenparameter unterschiedliche Ergebnisse. Es liegt nur im Fall der ES K5 eine Normalverteilung für die Personenfähigkeit vor, d. h., dass die Schülerinnen und Schüler weder über eine zu geringe noch über eine zu hohe Kompetenz verfügen. Bei der HE K5 liegt eine signifikant höhere mittlere Personenfähigkeit vor, die zur Verletzung der Normalverteilung für den Personenparameter führt. Der grafische Modelltest liefert sowohl bei der ES K5 als auch bei der HE K5 Hinweise für einen Unterschied in Bezug auf die Itemschwierigkeiten in Abhängigkeit von der zugrundeliegenden Personenfähigkeit, da sich zwischen den gebildeten Teilstichproben Differenzen bei den Itemparametern ausmachen lassen. Aufgrund der hohen erklärten Varianz sind die Unterschiede fast ausschließlich auf die Item- und Personenparamater zurückzuführen.

[23] Es ist anzumerken, dass aufgrund der großen Stichproben komplexere Modelle immer eine bessere Modellpassung aufweisen und sich signifikant von weniger komplexen Modellen unterscheiden (vgl. Hattie, 1984; Hattie, 1985).

8.1.3.2 EINDIMENSIONALE SKALIERUNGEN

Anhand der Ergebnisse der Ursprungsmodelle der eindimensionalen Skalierungen wird darauf eingegangen, ob die Stichprobengröße Einfluss auf die Verletzung statistischer Kriterien der Itemparameter zur Folge hat. Dazu sind in der Tabelle 8.15 die Anzahl und statistischen Auffälligkeiten der ganzen Wörter und Struktureinheiten für die ES und HE K5 vermerkt.

Tabelle 8.15: Anzahl und statistische Auffälligkeiten der ganzen Wörter und Struktureinheiten – ES & HE K5 – Eindimensionale Skalierungen, Ursprungsmodelle

	Ganzes Wort	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
	Wörter	Struktureinheiten				
Ursprungsmodell ES	54	51	49	24	71[24]	38
Item-Fit & Trennschärfe	-2	-	-1	-	-1	-2
Item-Fit	-2	-	-	-	-2	-
Trennschärfe	-1	-3	-6	-	-6	-3
Ursprungsmodell HE	54	51	49	24	72	38
Item-Fit & Trennschärfe	-2	-	-	-	-1	-1
Item-Fit	-	-	-	-	-1	-1
Trennschärfe	-2	-4	-10	-1	-16	-3

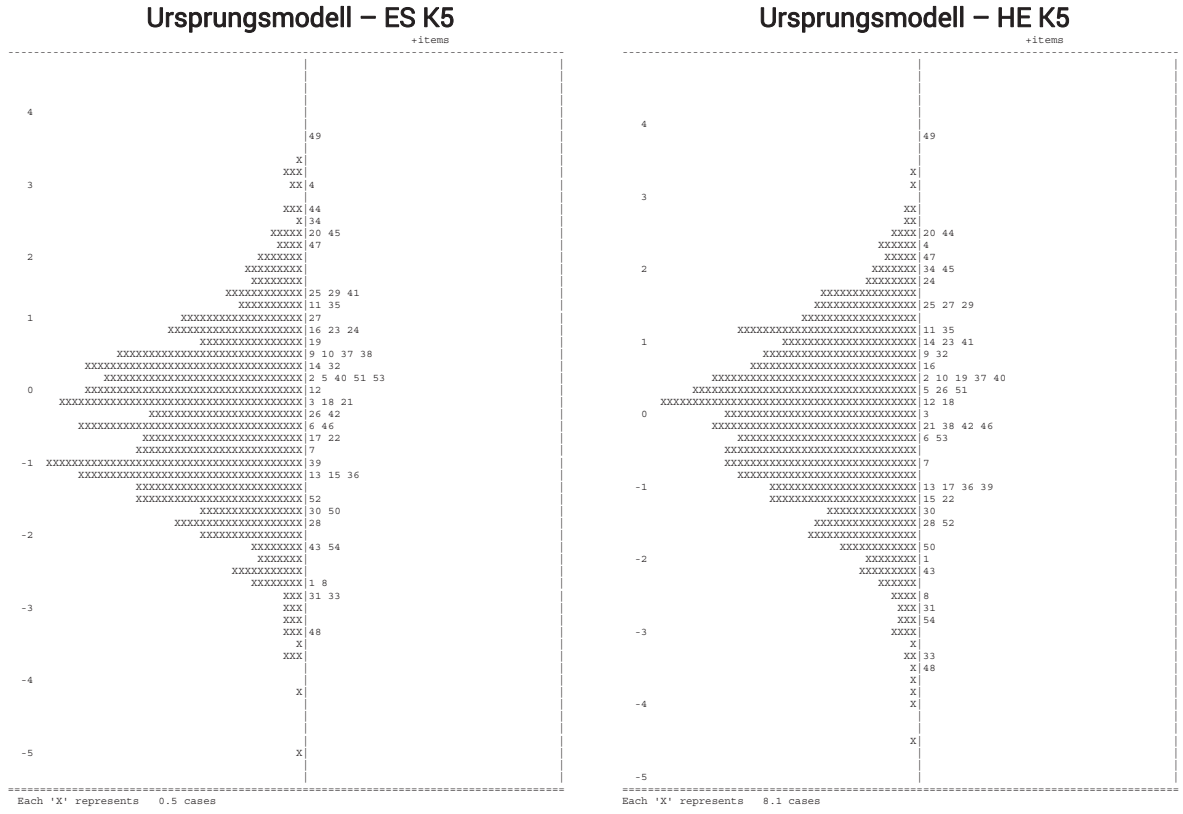
Die Anzahl der statistischen Auffälligkeiten verhält sich auf der Ebene der ganzen Wörter für die ES und HE nahezu identisch, da sich die beiden Studien in nur einem Wort unterscheiden. Dagegen bewirkt die größere Stichprobe der HE auf Ebene der Teilkompetenzen, dass mehr Struktureinheiten ausgeschlossen werden müssen. Ebenso ist zu erkennen, dass mit steigender Fallzahl die Verletzung des Kriteriums für die Trennschärfe zunimmt. Es lässt sich also

[24] Eine Struktureinheit wurde von allen Schülerinnen und Schülern richtig gelöst und wurde direkt ausgeschlossen.

folgern, dass der Item-Fit weniger sensibel auf die Stichprobengröße reagiert als die Trennschärfe.

Inwiefern sich die Skalierungsergebnisse im Detail auf eindimensionaler Ebene zwischen der ES und HE K5 unterscheiden, wird an dieser Stelle exemplarisch für das ganze Wort veranschaulicht, indem in Abbildung 8.21 jeweils die Ursprungsmodelle mit den Verteilungen der Item- und Personenparameter der beiden Studien miteinander verglichen werden. Inwiefern sich Unterschiede bei den Teilkompetenzen ergeben, wird nachfolgend kurz referiert.

Abbildung 8.21: Ganzes Wort im Vergleich – ES & HE K5 – Eindimensionale Skalierungen, Ursprungsmodelle



Hinsichtlich der Verteilung der Item- und Personenparameter ist bei der ES und HE K5 eine relativ große Ähnlichkeit zu erkennen. Die Itemparameter bieten eine gute Passung zur Verteilung der Personenfähigkeit, d. h., dass das Schwierigkeitsniveau des Testinstruments angemessen ist und die getesteten Wörter weder zu leicht noch zu schwer für die Schülerinnen und

Schüler sind. So konnten in beiden Studien anhand der ermittelten Personenfähigkeiten 25 leichte und 29 schwere Wörter identifiziert werden.

Die beiden Ursprungsmodelle der ES und HE K5 unterscheiden sich bezüglich der statistischen Auffälligkeiten im Fall von fünf Wörtern. Bei der ES K5 fallen drei Wörter wegen eines Item-Fits ≥ 1.20 (2 Wörter) bzw. einer Trennschärfe ≤ 0.25 (1 Wort) auf, die für die HE K5 bestehen bleiben. Umgekehrt umfasst die ES K5 zwei Wörter, die bei der HE K5 aufgrund einer zu geringen Trennschärfe entfallen. Insgesamt fallen die Abweichungen zu den Referenzwerten gering aus, d. h., dass die statistisch auffälligen Wörter zwischen 0.01 und 0.05 von den als zulässig definierten Werten abweichen. So verbleiben 50 von den ursprünglich 54 Wörtern bei der HE K5 nach Ausschluss der statistisch auffälligen Wörter, also ein Wort mehr als in der ES K5 mit 49 Wörtern. Demnach liegen in beiden Studien 49 identische Wörter vor, auf deren Itemparameter in Form des Item-Fits, der Trennschärfe sowie der Itemschwierigkeit (Kapitel 3.1.1) nachfolgend im Detail eingegangen wird.

In Bezug auf den Item-Fit zeigt sich, dass der Item-Fit bei der HE K5 im Vergleich zur ES K5 in 18 Fällen gesunken und in 31 Fällen gestiegen ist. Die Abweichungen bewegen sich zwischen -0.12 und 0.11 und sind als vertretbar zu bewerten, da es zu keiner Verletzung der Referenzwerte führt. Daraus ist abzuleiten, dass sich wegen der unterschiedlichen Stichprobengrößen größtenteils niedrigere Item-Fits bei kleinen Stichproben und höhere Item-Fits für größere Stichproben ergeben.

Die Trennschärfe der verbleibenden 49 ganzen Wörter sinkt bzw. steigt bei der HE K5 im Vergleich zur ES K5 bei 26 bzw. 23 Wörtern, wobei sich die Abweichungen der Trennschärfen zwischen -0.21 und 0.15 bewegen. Folglich kann die Trennschärfe der ganzen Wörter sowohl mit kleineren als auch größeren Stichproben zufriedenstellend ermittelt werden. Dennoch lässt sich bei der größeren Stichprobe der HE K5 erkennen, dass die Trennschärfe der ganzen Wörter tendenziell abnimmt bzw. auf die etwas höhere Personenfähigkeit der Schülerinnen und Schüler der HE K5 zurückzuführen ist.

Die Abweichungen der in Logits ermittelten Itemschwierigkeiten lassen sich nach Paek (2002, zitiert nach Wilson, 2005) einordnen, wodurch sich folgende Situation im Vergleich von ES K5 und HE K5 ergibt:

Tabelle 8.16: Differenzen der Itemschwierigkeiten – ES K5 & HE K5

logit-Differenz	Anzahl ganze Wörter
< 0.426 negligible	39
≥ 0.426 & ≤ 0.638 intermediate	6
≥ 0.638 large	4

Bezogen auf die 49 ganzen Wörter ist zu verzeichnen, dass bei der HE K5 25 ganze Wörter leichter bzw. 24 ganze Wörter schwerer werden. Die Vielzahl der Abweichungen bei den ermittelten Itemschwierigkeiten (≤ 0.638) ist nach Paek zu vernachlässigen. Lediglich sechs ganze Wörter weisen eine als mittlere Differenz zu bezeichnende Abweichung bei den Itemschwierigkeiten auf. Von diesen sechs ganzen Wörtern werden bei der HE K5 fünf Wörter leichter und ein Wort schwerer. Eine große Differenz konnte bei vier verbleibenden ganzen Wörtern festgestellt werden. Konkret werden jeweils zwei ganze Wörter bei der HE K5 im Vergleich zur ES K5 deutlich leichter bzw. schwerer. Das Verhältnis der leichten und schweren Wörter beträgt sowohl in der ES K5 als auch in der HE K5 22 zu 28[25] ganze Wörter. Dies deutet auf ein vergleichbares Schwierigkeitsniveau durch die eingesetzten Aufgaben hin.

Bei den Teilkompetenzen zeigt sich ebenfalls, dass je nach Studie unterschiedliche Struktureinheiten bestehen bleiben bzw. wegfallen. Es ergeben sich die in Tabelle 8.17 festgehaltenen Abweichungen.

[25] Exklusiv eines Items, das von einer leichten zu einer schweren Itemschwierigkeit wechselt.

Tabelle 8.17: Unterschiede der Struktureinheiten – ES & HE K5 – Eindimensionale Skalierung

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
	Struktureinheiten				
Ursprungsmodell ES	3	9	1	15	--
Nur in HE enthalten	1	3	0	3	--
Nur in ES enthalten	2	6	1	12	--
Ursprungsmodell HE	3	9	1	15	-
Nur in ES enthalten	2	6	1	12	--
Nur in HE enthalten	1	3	0	3	--

Demnach verbleiben bei der ES K5 auch Struktureinheiten, die bei der HE K5 wegfallen. Dieser Umstand ist durch das Zusammenspiel von der zugrundeliegenden Itemschwierigkeit und Personenfähigkeit in Abhängigkeit von der Stichprobengröße bedingt, da sich die Reduzierung der Struktureinheiten sichtlich häuft. Die Ergebnisse aus den vielfältigen statistischen Analysen lassen den Schluss zu, dass – ähnlich wie bei dem ganzen Wort – die Unterschiede auf der Ebene der fünf Teilkompetenzen vertretbar sind und sich keine stark bemerkbaren Differenzen zwischen den Studien ergeben.

Weitergehend sind in der nachfolgenden Tabelle 8.18 die deskriptiven Statistiken der Ursprungsmodelle für die ES und HE K5 für das ganze Wort und der Teilkompetenzen der eindimensionalen Skalierungen zusammengestellt.

Tabelle 8.18: Deskriptive Statistiken – ES & HE K5 – Eindimensionale Skalierungen, Ursprungsmodelle

	Ganzes Wort	Phono-graphisch-silbischer Kern-bereich	Morpho-logischer Kern-bereich	Peripherie-bereich	Prinzip der Wort-bildung	Wortüber-greifendes Prinzip
	Wörter	Struktureinheiten				
Ursprungsmodell ES						
MW	-0.38/0.00	2.41/0.00	1.80/0.00	0.86/0.00	1.83/0.00	2.05/0.00
SE	0.08/0.22	0.08/0.20	0.07/0.17	0.08/0.30	0.07/0.16	0.07/0.30
NV	✓/✓	✓/✓	✓/✓	✓/✓	✓/✓	✓/✓
Ursprungsmodell HE						
MW	-0.11/0.00	2.90/0.00	2.11/0.00	1.14/0.00	2.19/0.00	2.17/0.00
SE	0.02/0.22	0.02/0.18	0.02/0.19	0.02/0.30	0.01/0.17	0.02/0.29
NV	×/✓	×/✓	×/✓	×/✓	×/✓	×/✓

Angabe in Logits für die Personenfähigkeit/Itemschwierigkeit

Wie bereits aufgezeigt wurde, liegt bei der HE K5 eine höhere Personenfähigkeit vor, die beim ganzen Wort und bei den Teilkompetenzen zu erkennen ist. Aufgrund der Ergebnisse sind die Schülerinnen und Schüler der HE K5 bis auf das wortübergreifende Prinzip signifikant besser als die Schülerinnen und Schüler der ES K5. Dies beeinträchtigt jedoch nicht die Aussagekraft der Ergebnisse aus der Entwicklungsstudie für die Haupterhebung, da das Testinstrument trotz unterschiedlicher Itemschwierigkeiten zuverlässig misst.

Inwiefern sich Unterschiede zwischen der ES und HE K5 in Bezug auf die prozentuale Lösungshäufigkeit der ganzen Wörter und der Teilkompetenzen in Abhängigkeit von der Itemschwierigkeit und Personenfähigkeit ergeben, wird nun dargestellt.

Tabelle 8.19: Prozentuale Lösungshäufigkeiten – ES & HE K5

		Ganzes Wort	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
		Wörter	Struktureinheiten				
Ursprungsmodell ES	MW	44.26	83.22	77.65	62.99	77.68	78.49
	(SE)	(1.12)	(0.81)	(0.89)	(1.18)	(0.85)	(0.71)
Ursprungsmodell HE	MW	48.21	88.24	80.69	67.02	81.94	80.03
	(SE)	(0.27)	(0.15)	(0.20)	(0.29)	(0.16)	(0.17)

Angaben in Prozent

Die prozentualen Lösungshäufigkeiten zeigen die zu erwartenden Unterschiede zwischen den beiden Studien und bestätigen noch einmal die höhere Personenfähigkeit der Schülerinnen und Schüler der HE K5 auf der Ebene der ganzen Wörter und Teilkompetenzen.

Anhand der Reliabilitäten der eindimensionalen Skalierungen für das ganze Wort und die fünf Teilkompetenzen wird aufgezeigt, ob sich Unterschiede bei der Erfassung des differenziellen Kompetenzmodells mit dem sprachsystematischen Rechtschreibtest zwischen der ES und HE K5 ergeben.

Tabelle 8.20: Reliabilitäten – ES & HE K5 – Eindimensionale Skalierungen

	Ganzes Wort	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
	Wörter	Struktureinheiten				
Ursprungsmodell ES	0.93	0.86	0.87	0.84	0.90	0.79
Ursprungsmodell HE	0.92	0.79	0.84	0.85	0.84	0.78

In Anbetracht der unterschiedlichen Stichprobengrößen der beiden Studien wird bezüglich der ermittelten Reliabilitäten deutlich, dass die Erfassung des differenziellen Kompetenzmodells bei der ES im Gegensatz zur HE minimal besser erfolgt. Dennoch ist in beiden Studien die zu-

verlässige Erfassung der Rechtschreibkompetenz gewährleistet, da alle Reliabilitäten über dem Wert von 0.70 liegen (vgl. Moosbrugger & Kelava, 2012, S. 11).

8.1.3.3 MEHRDIMENSIONALE SKALIERUNG

Hinsichtlich der ermittelten Item-, Personen- und Modellparametern wird nun für die mehrdimensionale Skalierung aufgezeigt, ob sich die ES und HE K5 diesbezüglich unterscheiden.

Tabelle 8.21: Anzahl und statistische Auffälligkeiten der Struktureinheiten – ES & HE K5 – Mehrdimensionale Skalierung

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
	Struktureinheiten				
Ursprungsmodell ES	51	49	24	71[26]	38
Item-Fit & Trennschärfe	-	-2	-	-2	-2
Item-Fit	-	-	-	-	-
Trennschärfe	-3	-5	-1	-11	-7
Ursprungsmodell HE	51	49	24	72	38
Item-Fit & Trennschärfe	-13	-	-	-2	-1
Item-Fit	-	-	-1	-	-1
Trennschärfe	-	-9	-1	-20	-4

Wie bereits bei der eindimensionalen Skalierung festgestellt werden konnte, werden bei der HE K5 mehr Struktureinheiten von den Analysen ausgeschlossen, die vielfach das statistische Kriterium für die Trennschärfe verletzen. Infolgedessen ist festzuhalten, dass bereits mit den Entwicklungsstudien eine ausreichende Grundlage gegeben ist, um adäquat die Itemparameter zu ermitteln. Den einzigen merkbaren Gewinn bei einer großen Stichprobe stellt die Sensibilität für die Verletzung statistischer Grenzen dar, die sichtlich mit der Fallzahl ansteigt.

[26] Eine Struktureinheit wurde von allen Schülerinnen und Schülern richtig gelöst und wurde direkt ausgeschlossen.

Tabelle 8.22: Veränderungen der deskriptiven Statistiken – ES & HE K5 – Mehrdimensionale Skalierung

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
			Struktureinheiten		
Ursprungsmodell ES					
MW	2.40	1.80	0.85	1.83	2.03
SE	0.08	0.07	0.08	0.07	0.06
Ursprungsmodell HE					
MW	2.89	2.10	1.13	2.18	2.15
SE	0.02	0.02	0.02	0.01	0.02
Angabe in Logits für die Personenfähigkeit					

An dieser Stelle sind im Gegensatz zur eindimensionalen Skalierung keine neuen Erkenntnisse in Bezug auf den Vergleich der beiden Studien zu erwarten, da die deskriptiven Statistiken der ein- und mehrdimensionalen Skalierungen recht ähnlich sind. So zeigt sich auch hier, dass die Schülerfähigkeiten bis auf eine Ausnahme bei der HE K5 signifikant besser sind.

Im Folgenden wird der Fokus auf die latenten Korrelationen und Reliabilitäten des differenziellen Rechtschreibkompetenzmodells in Abbildung 8.22 gelenkt, um mögliche stichprobenbedingte Unterschiede aufzudecken.

Abbildung 8.22: Latente Korrelationen und Reliabilitäten der fünf Teilkompetenzen für die Ursprungsmodelle – ES & HE K5 – Mehrdimensionale Skalierung

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
Phonographisch-silbischer Kernbereich	0.94 0.92	0.98	0.96	0.97	0.91
Morphologischer Kernbereich	0.97	0.93 0.94	0.99	0.95	0.84
Peripheriebereich	0.96	0.99	0.92 0.93	0.95	0.83
Prinzip der Wortbildung	0.96	0.95	0.95	0.94 0.92	0.90
Wortübergreifendes Prinzip	0.81	0.78	0.79	0.83	0.86 0.82

Latente Interkorrelationen (dargestellt auf der oberen und unteren Dreiecksmatrix) sowie Reliabilitäten (EAP/PV) (dargestellt auf der Diagonalen) der fünf Teilfähigkeiten für das Ursprungsmodell (obere Angaben) in der ES K5 und das Ursprungsmodell (untere Angaben) in der HE K5.

Hinsichtlich der latenten Korrelationen zeigen sich sowohl für die ES K5 als auch für die HE K5 hohe Zusammenhänge zwischen den Teilkompetenzen, die von 0.78 bis 0.99 variieren. Nennenswerte Unterschiede lassen sich anhand der Korrelationskoeffizienten nicht ausmachen, wobei es kleine Unterschiede in der Stärke der Zusammenhänge gibt. So korrelieren die Teilkompetenzen im Fall der ES K5 überwiegend höher, aber es gibt dennoch keinen merklichen Unterschied in der Eigenständigkeit der Dimensionen, da die Korrelationen der HE K5 maximal um 0.10 geringer sind.

8.1.3.4 DIMENSIONALITÄT DES KOMPETENZKONSTRUKTS

Der Vergleich der Entwicklungsstudie (ES) und der Haupterhebung (HE) in Klassenstufe 5 (K5) schließt mit der Darstellung der Überprüfung der Dimensionalität des Kompetenzkonstrukts ab, die durch vier Modellgeltungstests erfolgt. Anzumerken ist hierbei, dass die ermittelten Statistiken nicht direkt für die jeweiligen Modelle verglichen werden können, da sie stichprobenabhängig sind. Vielmehr kann an dieser Stelle gezeigt werden, mit wie vielen Dimensionen

sich das differenzielle Kompetenzmodell anhand der empirischen Daten am besten beschreiben lässt.

Tabelle 8.23: Modellgeltungstests – ES & HE K5 – Mehrdimensionale Skalierung

	1D	2D	4D	5D
Deviance	46234.40556 665319.74747	46194.72762 662704.11856	46131.86584 661618.79388	**46116.82433** **661366.40395**
AIC	46636.40556 665685.74747	46640.40556 663074.11856	46551.86584 662002.79388	**46546.82433** **661760.40395**
CAIC	47580.52135 667060.99078	47593.91554 664464.39185	**47538.25547** 663445.67211	47556.69943 **663240.85713**
BIC	47379.52135 666877.99078	47390.91554 664279.39185	**47328.25547** 663253.67211	47341.69943 **663043.85713**

ES K5 (N = 298, 200 Struktureinheiten)
HE K5 (N = 4.989, 181 Struktureinheiten)

Die Ergebnisse der ES K5 sprechen für ein vier- oder fünfdimensionales Modell, um die fünf Teilkompetenzen zu modellieren. Dagegen zeigt sich bei der HE K5, dass das fünfdimensionale Modell vorzuziehen ist. Mit den bereits in Kapitel 8.1.2.4 erwähnten Einschränkungen ist in diesem Fall davon auszugehen, dass die große Stichprobe der HE K5 ein differenzierteres Modell begünstigt. Dies führt zu dem Schluss, dass die Größe der Stichprobe eher hinderlich für die Überprüfung der Dimensionalität sein kann und aus didaktischer Sicht die Entwicklungsstudien zur Untersuchung der Dimensionalität des Kompetenzkonstrukts geeigneter erscheinen. Da sich im Rahmen dieser Arbeit kein eindeutiges Muster zeigt und es deutliche Tendenzen aufgrund der vorrangegangenen Korrelationsanalyse in Richtung einer Veränderung bzw. Reduzierung des fünfdimensionalen Modells gibt, wird in diese Richtung weitergeforscht (vgl. Blatt, Frahm, Prosch, Jarsinski & Voss, in Vorb.).

8.2 ERGEBNISSE ZUR KOMPETENZENTWICKLUNG

Weiterführend steht in diesem Kapitel die längsschnittliche Erfassung der Kompetenzentwicklung im Fokus, wobei es um die Beantwortung der zweiten Forschungsfrage geht:

F2: Mit welchen methodischen Verfahren lässt sich die Entwicklung der Rechtschreibkompetenz am verlässlichsten erfassen?

Konkret werden drei Skalierungsverfahren zur längsschnittlichen Modellierung der Rechtschreibkompetenz auf der Ebene des ganzen Wortes verglichen, um

- die Entwicklung der Rechtschreibkompetenz von der Klassenstufe 6 bis zur Klassenstufe 7 adäquat erfassen zu können.
- die Nutzung von Ankeritems und zeitpunkspezifischen Items zur längsschnittlichen Erfassung der Rechtschreibung zu erproben, indem unterschiedliche Skalierungs- und Fixierungsvarianten verglichen werden.
- ein ideales Verfahren für die Modellierung des differenziellen Rechtschreibkompetenzmodells mit den fünf Teilkompetenzen zu finden.

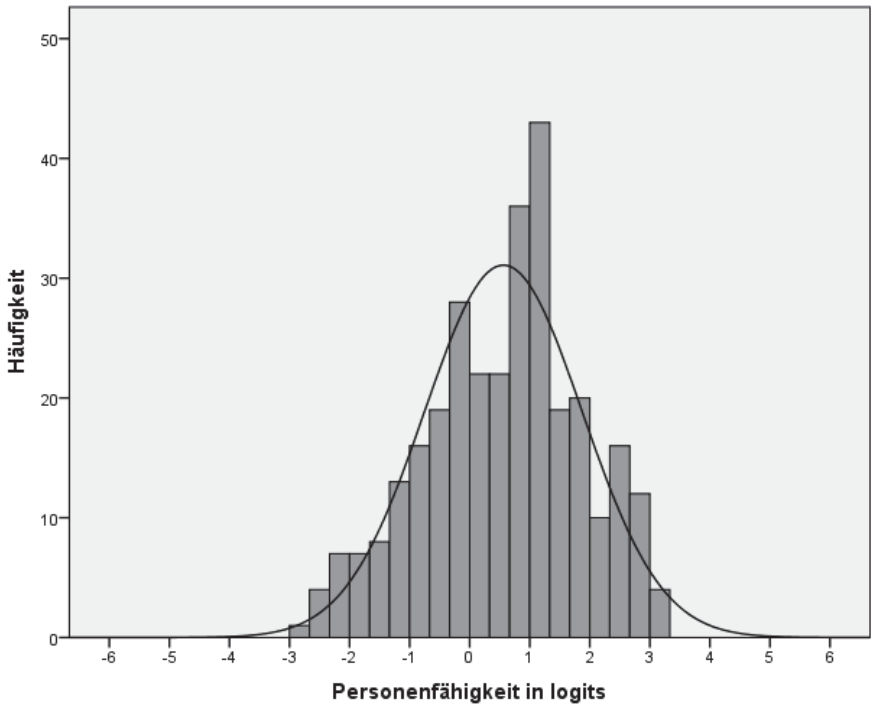
8.2.1 DATENGRUNDLAGE DER ENTWICKLUNGSSTUDIEN K6 UND K7

Die längsschnittlichen Daten der Entwicklungsstudien (ES) in Klassenstufe 6 (K6) und 7 (K7) werden auf der Ebene des ganzen Wortes in einem ersten Schritt durch die Überprüfung einer Normalverteilung und der Annahme des Rasch-Modells untersucht. Es werden weitergehend die längsschnittliche Modellierung und Entwicklung der Rechtschreibkompetenz auf der Grundlage eindimensionaler Skalierungen des ganzen Wortes aufgezeigt. Dies wird mit den drei Verfahren zur längsschnittlichen Kompetenzmodellierung von personenspezifischen Veränderungen, die in Kapitel 3.3.1 vorgestellt wurden, umgesetzt, und zwar mit der getrennten Skalierung, der Skalierung mit virtuellen Personen und der Skalierung mit latenten Dimensionen. Die einzelnen Verfahren werden hinsichtlich ihrer Möglichkeiten und Grenzen miteinander verglichen. Ebenso wird die Erfassung der Kompetenzentwicklung mithilfe von Ankeritems und zeitpunktspezifischen Items untersucht. Mit den Analysen wird das Ziel verfolgt, für die längsschnittliche Modellierung des differenziellen Rechtschreibkompetenzmodells und für die Untersuchung der Kompetenzstruktur, um die es im nächsten Kapitel 8.3 geht, ein verlässliches und effizientes Verfahren für die Skalierung zu erforschen.

8.2.1.1 DATENPRÜFUNG

Zur Überprüfung einer Normalverteilung für die Personenparameter der ES K6 und K7 wird der Kolmogorov-Smirnov-Test auf der Ebene des ganzen Wortes vorgestellt. Die jeweiligen Personenfähigkeiten in Logits sind in Abbildung 8.23 beispielhaft für die ES K6 als Histogramm gemeinsam mit der Normalverteilungskurve dargestellt. Die Ergebnisse für die ES K7 werden referiert und die dazugehörigen Abbildungen finden sich im Anhang (vgl. Abbildung 11.1).

Abbildung 8.23: Verteilung der Personenfähigkeit für das ganze Wort – ES K6 (N = 307)

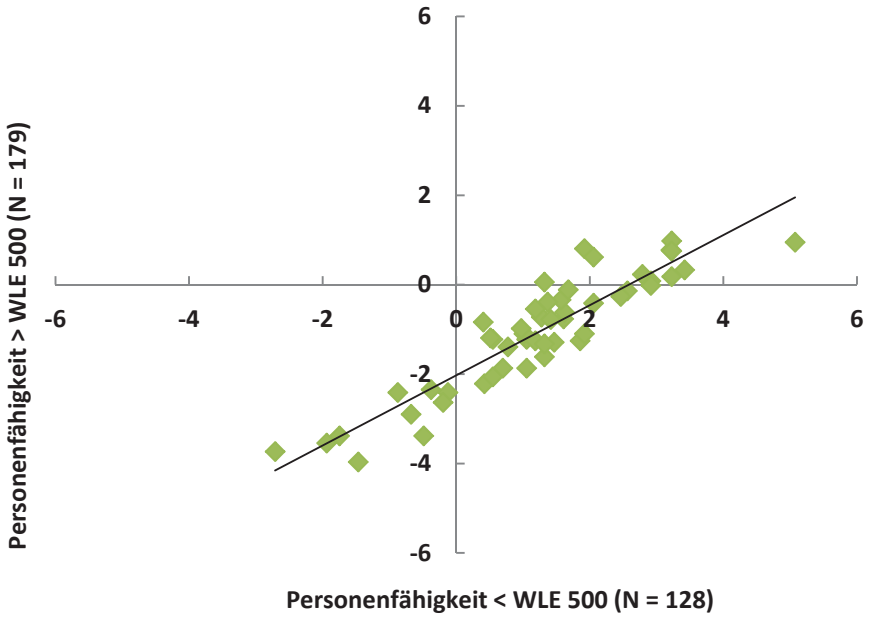


Die Personenfähigkeit der ES K6 weist einen Mittelwert und eine Standardabweichung von 0.56 bzw. 1.31 Logits auf. Mit einem Signifikanzniveau von 0.30 kann die Normalverteilung für die Personenparameter nachgewiesen werden.

Für die ES K7 liegt keine Normalverteilung der Personenparameter (Sig. 0.05) bei einem Mittelwert und einer Standardabweichung von 0.64 bzw. 1.33 Logits vor (vgl. Abbildung 11.1 im Anhang)

Zur Annahme des Rasch-Modells werden die grafischen Modelltests für die ES K6 und ES mit den Itemschwierigkeiten der 50 Ankeritems[27] betrachtet, um mögliche Unterschiede bei den zwei Teilstichproben bestehend aus Schülerinnen und Schülern mit einer Personenfähigkeit unter und über einem WLE-Wert von 500 zu identifizieren. Abbildung 8.24 stellt beispielhaft für die ES K6 auf den beiden Achsen die Itemschwierigkeiten der Teilstichproben dar.

Abbildung 8.24: Grafischer Modelltest auf Basis der Itemschwierigkeiten – ES K6



Bei den Itemschwierigkeiten für die ES K5 zeigen sich bei den beiden Teilstichproben Differenzen zwischen 1.02 und 4.13 Logits. Die damit erklärte Varianz (R^2) beträgt 0.96. Im Fall der ES K7 zeigt sich ein ähnliches Bild, wobei die Differenzen (Minimum: 1.03; Maximum 3.55) und die erklärte Varianz (R^2 = 0.95) leicht zurückgehen (vgl. Abbildung 11.2 im Anhang).

Zusammenfassend zeigt sich, dass eine Normalverteilung der Personenfähigkeit nur bei der ES K6 vorliegt, wobei die Schülerinnen und Schüler aus der ES K7 im Vergleich zu den Schülerinnen und Schülern aus der ES K6 nicht signifikant besser sind. Aufgrund der Differenzen

[27] Die eigentlich 51 Ankeritems umfassen 20 Items des Längsschnitts von Klasse 5 bis 7 und nochmals 31 Items aus dem Längsschnitt von Klasse 6 bis 7, wobei ein Item aus dem ersten Längsschnitt wegen einer Konstanten für die Gruppe der Schülerinnen und Schüler mit einer Personenfähigkeit über dem WLE-Wert von 500 von dem gesamten Vergleich ausgeschlossen werden musste.

beim deskriptiven grafischen Modelltest lässt sich ableiten, dass es ohne signifikanzanalytische Prüfung Hinweise für die Verletzung der Personenhomogenität gibt und die Annahme einer unterschiedlichen Kompetenzstruktur bei Schülerinnen und Schülern mit einer hohen bzw. niedrigen Personenfähigkeit naheliegt. Um tiefergehende Statistiken zur Modellierbarkeit der Rechtschreibkompetenz zu erhalten, wird das Rasch-Modell für die ES K6 und K7 mit den erkennbaren Differenzen bei den Itemschwierigkeiten vorläufig angenommen (vgl. Kapitel 8.1.1.1 und 8.1.2.1).

Zur Einordnung der nachfolgenden Ergebnisse zur längsschnittlichen Kompetenzmodellierung wird die prozentuale Lösungshäufigkeit für die beiden Studien angeben.

8.2.1.2 PROZENTUALE LÖSUNGSHÄUFIGKEIT

Zur Einordung der nachfolgenden Ergebnisse bezüglich der Entwicklung der Personenfähigkeit werden einleitend die prozentuale Lösungshäufigkeit aller 51 Ankeritems sowie die Kombination aus den 51 Ankeritems und den zeitpunktspezifischen Items der ES K6 und ES K7 betrachtet.

Tabelle 8.24: Prozentuale Lösungshäufigkeit – ES K6 & K7

		51 Ankeritems	+ zeitpunktspezifische Items
ES K6	MW(SD)	53.04 (22.06)	58.40 (20.47)
ES K7	MW(SD)	57.04 (22.71)	63.05 (21.78)
Differenz		4.00	4.65
Effektstärke	(Cohen)	d = -0.18	d = -0.22

Demnach sind in der ES K6 53,04 Prozent und in der ES K7 57,04 Prozent der 51 Ankeritems von den Schülerinnen und Schülern korrekt gelöst worden. Dies entspricht einem Zuwachs von vier Prozent zwischen den Messzeitpunkten bzw. einem Effekt in Höhe von d = -0.18. Nimmt man die zeitpunktspezifischen Items hinzu, zeigt sich eine vergleichbare prozentuale Lösungshäufigkeit, wobei dies wegen der unterschiedlichen Anzahl an Items nur bedingt aussagekräftig ist. Im Fall der ES K6 mit insgesamt 75 Items beläuft sich die mittlere Lösungshäufigkeit auf 58,40 Prozent und bei der ES K7 sind 63,05 Prozent der insgesamt 92 Items von

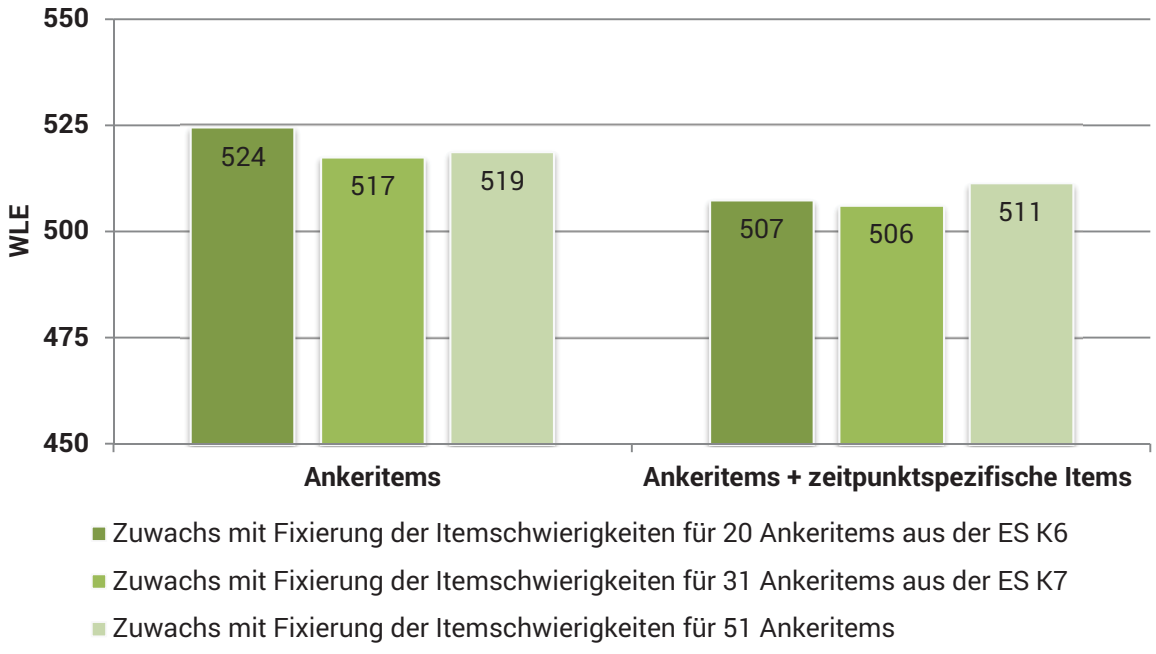
den Schülerinnen und Schüler korrekt gelöst worden. Dies entspricht einer Differenz der prozentualen Lösungshäufigkeit von 4.65 Prozent mit einer Effektstärke in Höhe von d = -0.22. Insgesamt sind die ermittelten Zuwächse auf der Ebene der prozentualen Lösungshäufigkeit als gering zu bewerten.

Die Ergebnisse der drei Verfahren zur längsschnittlichen Kompetenzmodellierung werden nun anhand der in Kapitel 7.1 beschriebenen Darstellungsform berichtet.

8.2.1.3 GETRENNTE SKALIERUNG

Im Fall einer getrennten Skalierung pro Messzeitpunkt ergibt sich, wie die Abbildung 8.25 veranschaulicht, eine mittlere Entwicklung zwischen 6 und 24 WLE-Punkten von Klassenstufe 6 nach 7. Die Unterschiede hängen zum einen davon ab, welche Skalierungsvariante gewählt wird: eine Skalierung mit Ankeritems bzw. mit Ankeritems und zusätzlichen zeitpunktspezifischen Items (vgl. Tabelle 7.5). Zum anderen hängen sie davon ab, welche Fixierungsvariante mit den Ankeritems der ES K6 und K7 verwendet werden.

Abbildung 8.25: Kompetenzentwicklung anhand einer getrennten Skalierung – ES K6 & K7

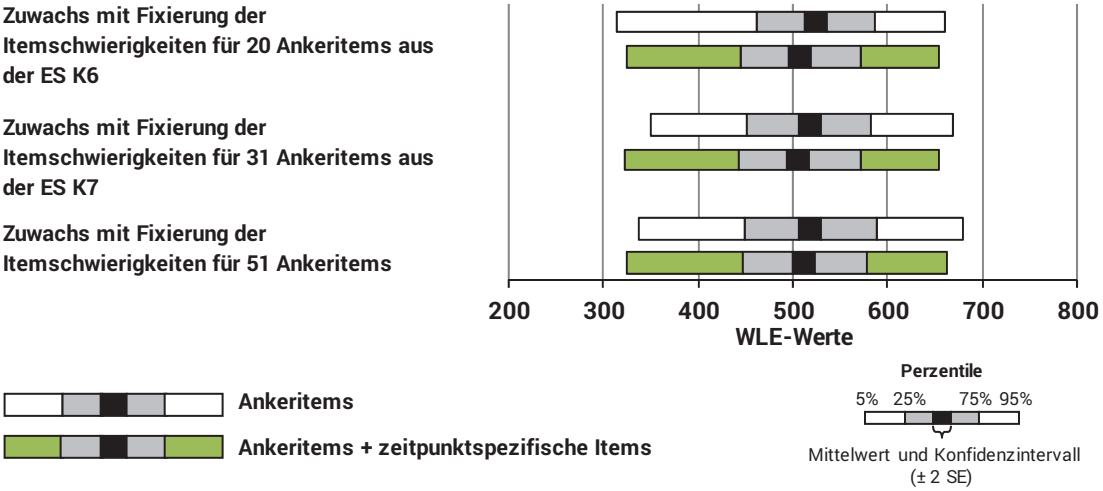


Bei der ausschließlichen Berücksichtigung der Ankeritems zur Ermittlung der Kompetenzentwicklung ergeben sich Zuwächse zwischen 17 und 24 WLE-Punkten. Auf der alleinigen Grundlage der 20 Ankeritems aus der ES K6 konnte die größte Entwicklung mit 24 WLE-Punkten (d = -0.24) festgestellt werden. Das lässt die Annahme zu, dass die Ankeritems aus der ES K6 zum Zeitpunkt der Messung in der ES K7 besser beherrscht werden, da die Lerninhalte bis dahin weitestgehend gefestigt sind und daher die Kompetenzentwicklung am größten ausfällt. Dagegen beläuft sich die Entwicklung basierend auf den 31 Ankeritems aus der ES K7 auf 517 WLE-Punkten (d = -0.17). Dies unterstützt die Vermutung, dass die Kompetenzentwicklung abhängig von der Bekanntheit der Lerninhalte ist, weswegen die Entwicklung in Abhängigkeit der Ankeritems aus der ES K6 in der ES K7 geringer ausfällt. Bei der Kombination der 20 bzw. 31 Ankeritems aus der ES K6 und ES K7 beträgt die ermittelte Personenfähigkeit 519 WLE-Punkte (d = -0.18), das in etwa der gemittelten Kompetenzentwicklung der beiden getrennten Fixierungsvarianten gleichkommt.

Werden neben den Ankeritems noch zusätzlich zeitpunktspezifische Items berücksichtigt, 24 Items in der ES K6 und 41 Items in der ES K7, fällt die ermittelte Kompetenzentwicklung niedriger aus. Liegt eine Fixierung der Itemschwierigkeiten bei 20 bzw. 31 Ankeritems vor, entwickelt sich die Personenfähigkeit um 7 bzw. 6 WLE-Punkte. Diese Veränderung entspricht einem zu vernachlässigenden bzw. sehr kleinen Effekt in Höhe von d = -0.07 und d = -0.06. Mit allen 51 Ankeritems wird eine Entwicklung der Personenfähigkeit in Höhe von 11 WLE-Punkten ermittelt, die einem sehr kleinen Effekt in Höhe von d = -0.11 entspricht.

Inwiefern Unterschiede bei der Ermittlung der Kompetenzentwicklungen für die getroffene Differenzierung bestehen, lässt sich grafisch anhand von Perzentilbändern durch das Zusammenspiel aus Mittelwert (MW), Standardfehler (SE) und Standardabweichung (SD) aufzeigen.

Abbildung 8.26: Vergleich der Skalierungsvarianten zur getrennten Skalierung – ES K6 & K7

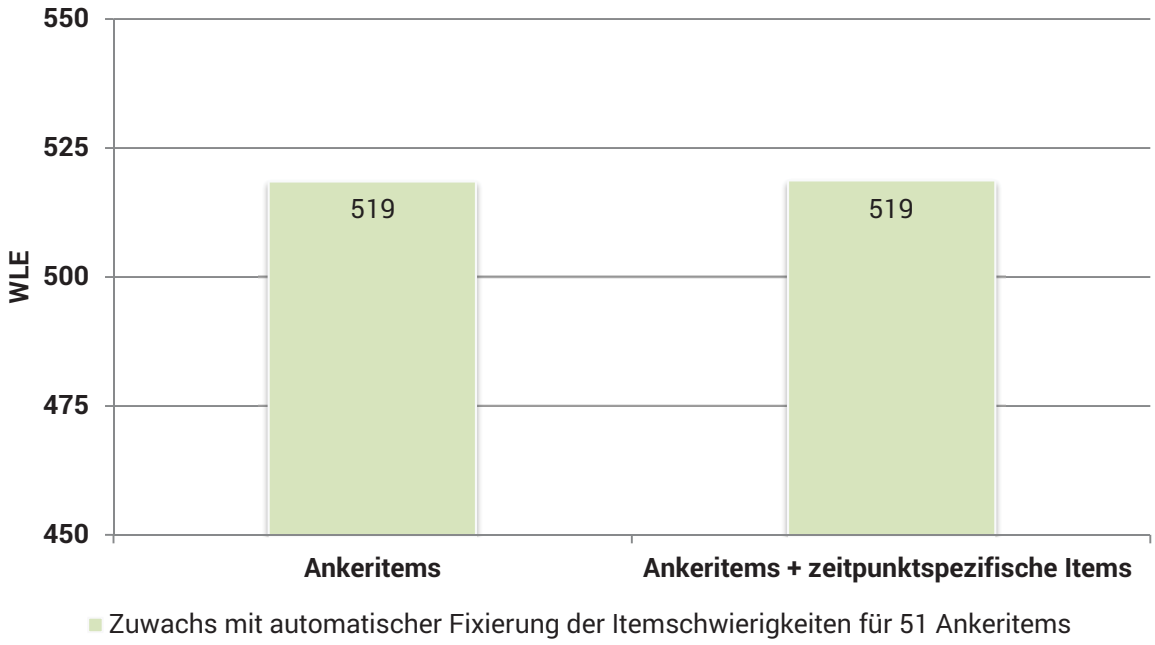


Mithilfe der oben dargestellten Perzentile für die unterschiedliche Anzahl an Ankeritems ist zu erkennen, dass keine signifikanten Unterschiede zwischen der Skalierung auf Basis der Ankeritems sowie der Skalierung mit zusätzlichen zeitpunktspezifischen Items bestehen, da sich die jeweiligen Konfidenzintervalle überlappen.

8.2.1.4 SKALIERUNG MIT VIRTUELLEN PERSONEN

Die Skalierung der beiden Messzeitpunkte als virtuelle Personen basiert auf einer automatischen Fixierung der Itemschwierigkeiten für die 51 Ankeritems, weswegen ein Vergleich der Fixierungsvarianten nicht möglich ist. Es konnte die in Abbildung 8.27 dargestellte Entwicklung der Personenfähigkeit ermittelt werden.

Abbildung 8.27: Kompetenzentwicklung anhand einer Skalierung mit virtuellen Personen – ES K6 & K7

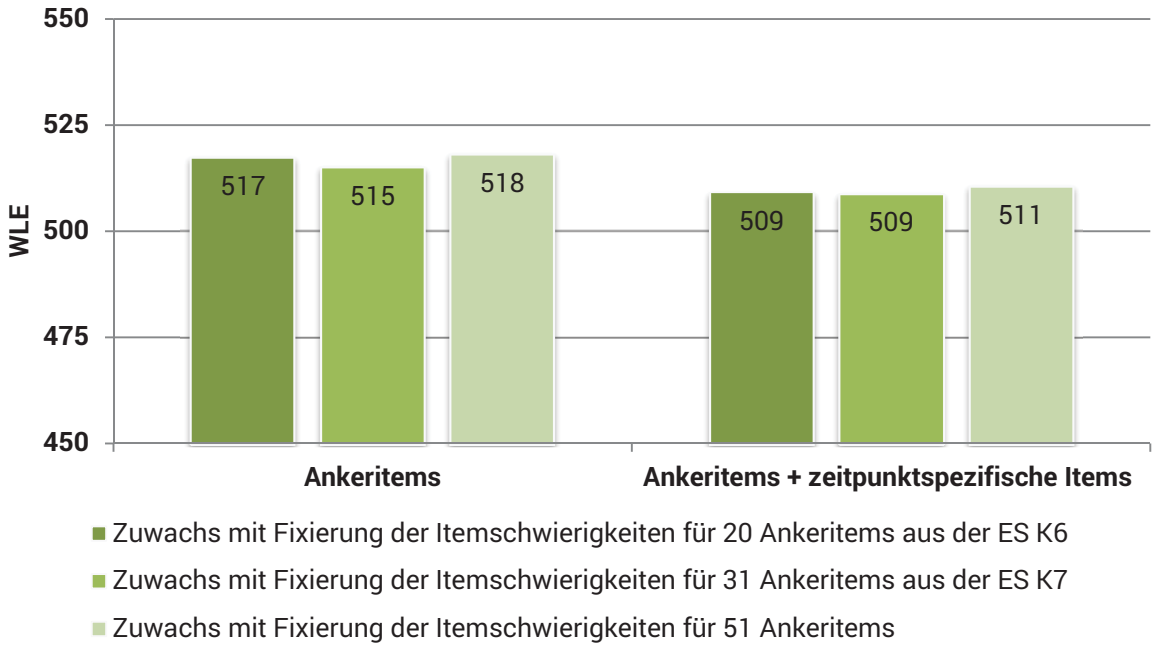


Demnach entwickelte sich die Personenfähigkeit der Schülerinnen und Schüler im Fall der längsschnittlichen Kompetenzmodellierung mit virtuellen Personen bei beiden Skalierungsvarianten um 19 WLE-Punkte. Dies entspricht einem annähernd kleinen Effekt von d = -0.18. Da es keine Unterschiede zwischen den Skalierungsvarianten gibt, stellt die Skalierung mit virtuellen Personen möglicherweise eine Alternative dazu dar, die längsschnittliche Kompetenzmodellierung mit zusätzlichen zeitpunktspezifischen Items durchzuführen. Auf die dadurch auftretenden Fragen wird abschließend in der Zusammenfassung der Skalierungsverfahren nochmals eingegangen.

8.2.1.5 SKALIERUNG MIT LATENTEN DIMENSIONEN

Das dritte Verfahren zur längsschnittlichen Kompetenzmodellierung bildet die zu den verschiedenen Messzeitpunkten erhobenen Schülerleistungen als latente Dimensionen in einem Modell ab, wodurch ein direkter Bezug zwischen den Messungen hergestellt wird. Mit diesem Verfahren konnte eine Entwicklung der Personenfähigkeit zwischen 9 und 18 WLE-Punkten ermittelt werden.

Abbildung 8.28: Kompetenzentwicklung anhand einer Skalierung mit latenten Dimensionen – ES K6 & K7

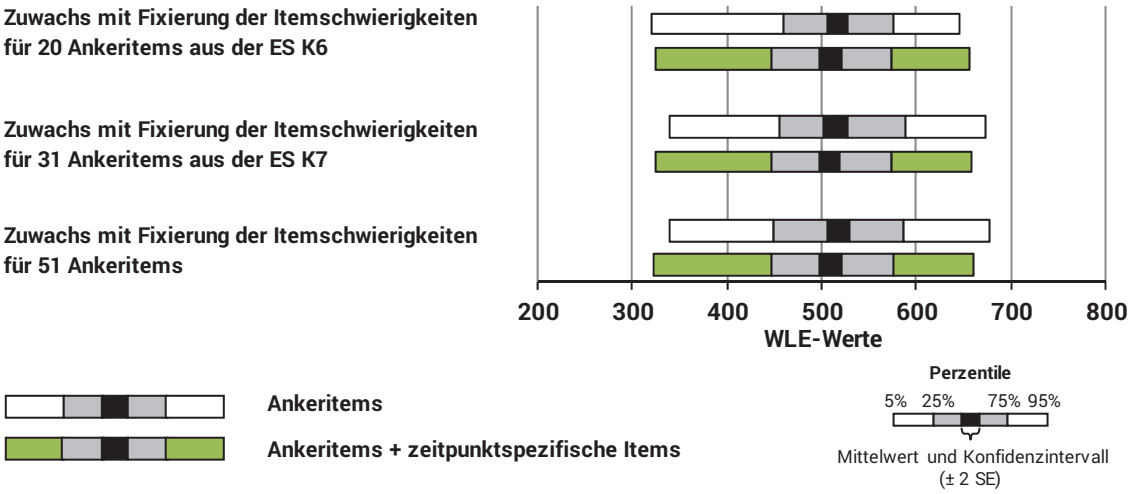


Werden ausschließlich die Ankeritems zur Ermittlung der Kompetenzentwicklung berücksichtig, ergeben sich Zuwächse zwischen 15 und 18 WLE-Punkten. Der kleinste Zuwachs konnte auf Basis der 31 Ankeritems aus der ES K7 mit 15 WLE-Punkten (d = -0.15) festgestellt werden. Erwartungsgemäß konnte mit den 20 Ankeritems aus der ES K6 eine höhere Entwicklung mit 517 WLE-Punkten (d = -0.18) beobachtet werden. Zusammengefasst beträgt die ermittelte Personenfähigkeit auf der Grundlage von 51 Ankeritems 518 WLE-Punkte (d = -0.18).

In Kombination mit zusätzlichen zeitpunktspezifischen Items fällt die ermittelte Kompetenzentwicklung erneut niedriger aus. So entspricht der Zuwachs bei den jeweiligen 20 bzw. 31 Ankeritems der ES K6 und ES K7 je 9 WLE-Punkte bzw. einem Effekt in Höhe von d = -0.09. Mit allen 51 Ankeritems ist die erreichte Entwicklung leicht höher und liegt bei 511 WLE-Punkten (d = -0.18).

Inwiefern Unterschiede bei der Ermittlung der Kompetenzentwicklungen bestehen, wird in Abbildung 8.29 ebenfalls mit Perzentilbändern veranschaulicht.

Abbildung 8.29: Vergleich der Skalierungsvarianten zur Skalierung mit latenten Dimensionen – ES K6 & K7



Demnach zeigen sich auch an dieser Stelle keine signifikanten Unterschiede zwischen der Skalierung auf Basis der Ankeritems sowie der Skalierung mit zusätzlichen zeitpunktspezifischen Items in Bezug auf die unterschiedliche Berücksichtigung der Anzahl von Ankeritems im Fall der Skalierung mit latenten Dimensionen.

8.2.1.6 BEWERTUNG DER SKALIERUNGSVERFAHREN

Die drei vorgestellten Verfahren zur längsschnittlichen Modellierung der Rechtschreibkompetenz zeigen auf der Ebene des ganzen Wortes je nach Skalierungs- und Fixierungsvariante kleinere Unterschiede bezüglich der ermittelten Personenfähigkeit. D. h., dass es auf der einen Seite zu unterschiedlichen Ergebnissen führt, wenn lediglich die Ankeritems bzw. zusätzlich auch die zeitpunktspezifischen Items skaliert werden. Auf der anderen Seite resultieren Unterschiede abhängig davon, welche Ankeritems zur Ermittlung der Kompetenzentwicklung berücksichtigt werden.

Die Ergebnisse der einzelnen Skalierungen sind in Tabelle 8.25 zusammengetragen. Sie werden auf Grundlage der prozentualen Lösungshäufigkeit interpretiert, die einen Effekt für die Entwicklung der Personenfähigkeit von der ES K6 zur ES K7 im Rahmen von d = -0.18 bis d = -0.22 erwarten lässt.

Tabelle 8.25: Gegenüberstellung der längsschnittlichen Skalierungsverfahren – ES K6 & K7

	Getrennte Skalierung			Virtuelle Personen			Latente Dimensionen		
Ankeritems									
	MW (SD|SE)	Dif.	d	MW (SD|SE)	Dif.	d	MW (SD|SE)	Dif.	d
Variante 1	524 (102|5.8)	24	-0.24	-	-	-	517 (097|6.0)	17	-0.18
Variante 2	517 (102|5.8)	17	-0.17	-	-	-	515 (105|6.0)	15	-0.15
Variante 3	**519 (105|6.0)**	**19**	**-0.18**	**519 (101|6.0)**	**19**	**-0.18**	**518 (104|6.0)**	**18**	**-0.18**
Ankeritems + zeitpunktspezifische Items									
	MW (SD|SE)	Dif.	d	MW (SD|SE)	Dif.	d	MW (SD|SE)	Dif.	d
Variante 1	507 (099|5.7)	7	-0.07	-	-	-	509 (100|6.0)	9	-0.09
Variante 2	506 (100|5.7)	6	-0.06	-	-	-	509 (100|6.0)	9	-0.09
Variante 3	**511 (101|5.8)**	**11**	**-0.11**	**519 (101|5.8)**	**19**	**-0.18**	**511 (101|6.0)**	**11**	**-0.11**

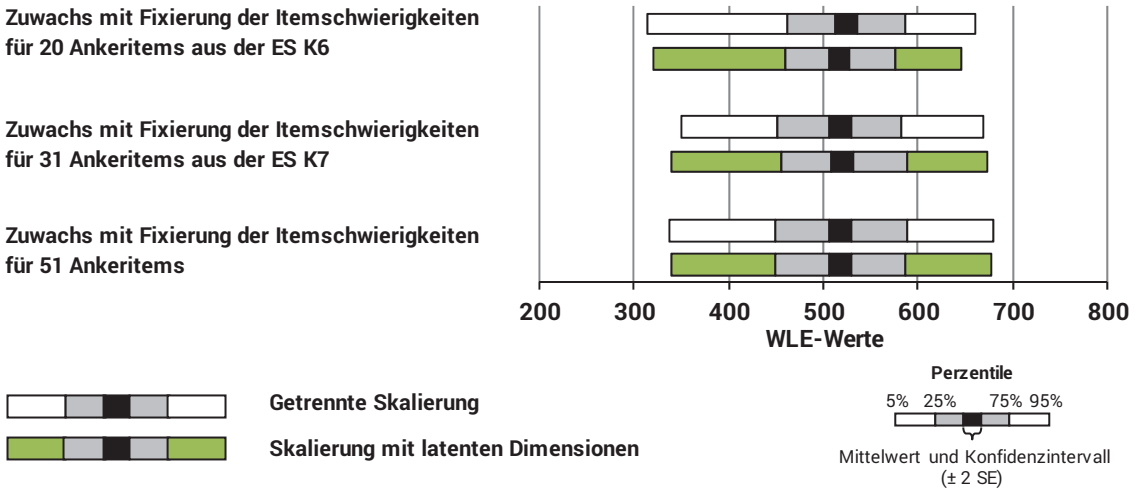
ES K7 (N = 307)
Variante 1: Zuwachs mit Fixierung der Itemschwierigkeiten für 20 Ankeritems aus der ES K6
Variante 2: Zuwachs mit Fixierung der Itemschwierigkeiten für 31 Ankeritems aus der ES K7
Variante 3: Zuwachs mit Fixierung der Itemschwierigkeiten für 51 Ankeritems

Betrachtet man nur die dritte Fixierungsvariante mit allen 51 Ankeritems, zeigt sich bei der Skalierungsvariante, die nur die Ankeritems berücksichtigt, ein Zuwachs von 18 oder 19 WLE-Punkten, was einem Effekt von d = -0.18 entspricht. Werden dagegen auch die zeitpunktspezifischen Items aufgenommen, zeigt sich mit Ausnahme bei der Skalierung mit virtuellen Personen eine Entwicklung der Personenfähigkeit in Höhe von 11 WLE-Punkten mit einem Effekt von d = -0.11. Aufgrund der automatischen Fixierung der Itemschwierigkeiten bei der Skalierung mit virtuellen Personen kann kein Unterschied zwischen den Skalierungsvarianten festgestellt werden, wobei die Personenfähigkeit zu je 19 WLE-Punkten zunimmt. Zusammenfassend lässt sich folgern, dass die Skalierungsvariante auf der alleinigen Grundlage der 51 Ankeritems mit den Ergebnissen zur prozentualen Lösungshäufigkeit übereinstimmt (vgl. Tabelle 8.4). Dagegen führen die zeitpunktspezifischen Items zu einer leichten und nicht signifikanten Verringerung der Kompetenzentwicklung. Aufgrund der nicht signifikanten Unterschiede kann die Veränderung bzw. Entwicklung der Personenfähigkeit sowohl auf Basis der Ankeritems als auch auf Basis aller Items in Form der Ankeritems und zeitpunktspezifischen Items bestimmt werden. Die Hinzunahme der zeitpunktspezifischen Items bietet aus didaktischer Perspektive

einen tiefergehenden Einblick bzw. zusätzliche Informationen über die Kompetenzentwicklung der Schülerinnen und Schüler in Abhängigkeit von unterschiedlichen Lerninhalten.

Insbesondere die getrennte Skalierung und die Skalierung mit latenten Dimensionen eignen sich zur längsschnittlichen Kompetenzmodellierung, da sie den Zuwachs im Vergleich der prozentualen Lösungshäufigkeit exakt bestimmen und die manuelle Fixierung der Itemschwierigkeiten bei der Kompetenzmodellierung ermöglichen. Wie sich die ermittelten Kompetenzentwicklungen bei den beiden Skalierungsverfahren zueinander verhalten, wird nachfolgend anhand von Perzentilbändern für die Skalierungsvariante mit den Ankeritems überprüft. Die Skalierungsvariante mit zusätzlichen zeitpunktspezifischen Items ist unter Abbildung 11.3 im Anhang zu finden.

Abbildung 8.30: Vergleich der Skalierungsverfahren mit Ankeritems – ES K6 & K7



Vergleicht man die einzelnen Konfidenzintervalle der beiden Skalierungsvarianten, so liegen in allen Fällen Überschneidungen vor, weswegen die getrennte Skalierung und die Skalierung mit latenten Dimensionen gleichermaßen zur längsschnittlichen Kompetenzermittlung geeignet sind.

Demnach lässt sich aufgrund der in diesem Kapitel dargestellten Ergebnisse folgern, dass die längsschnittliche Kompetenzmodellierung je nach Forschungsziel auf der Grundlage der Ankeritems bzw. mit zusätzlichen zeitpunktspezifischen Items zu verlässlichen Ergebnissen führt. Als Verfahren eignen sich die getrennte Skalierung und die Skalierung mit latenten Di-

mensionen, da sie zu übereinstimmenden Ergebnissen führen und je nach Verfahren spezifische Vorteile bieten. Die getrennte Skalierung und die Skalierung mit latenten Dimensionen bieten identische Möglichkeiten zur Modellanpassung, z. B. die manuelle Fixierung der Itemschwierigkeiten. Darüber hinaus stellt die Skalierung mit latenten Dimensionen im Gegensatz zur getrennten Skalierung einen direkten Bezug zwischen den Messzeitpunkten her, indem die gesamten Item- und Personenparameter sowie die latenten Korrelationen der beiden Messungen in einem gemeinsamen Modell bestimmt werden.

Folglich liegt für die differenzielle Modellierung des Rechtschreibkompetenzmodells und der Analyse der Kompetenzstruktur, die im nächsten Kapitel 8.3 dargestellt wird, der Schluss nahe, dass die längsschnittliche Modellierung des differenziellen Rechtschreibkompetenzmodells mithilfe von getrennten Skalierungen mit latenten Dimensionen den größtmöglichen Nutzen erzielt, d. h., dass getrennte Skalierungen und Skalierungen mit latenten Dimensionen kombiniert werden. Dies begründet sich damit, dass durch die getrennte Skalierung mit latenten Dimensionen für die längsschnittliche Kompetenzmodellierung eine Optimierung des Skalierungsaufwands erzielt wird, da lediglich zwei getrennte Skalierungen mit latenten Dimensionen für die Modellierung der fünf Teilkompetenzen über zwei Messzeitpunkte notwendig sind. Zudem liefert dieses Vorgehen den Vorteil, dass mit den getrennten Skalierungen mit latenten Dimensionen für zwei Messzeitpunkte die Zusammenhänge sowie die Reliabilitäten der einzelnen Dimensionen bzw. Teilkompetenzen direkt bestimmt werden.

8.3 ERGEBNISSE ZUR KOMPETENZSTRUKTUR

Um sich tiefergehend mit der erfassten Kompetenz zu befassen, wird mit der dritten Forschungsfrage das differenzielle Kompetenzmodell längsschnittlich untersucht und die fünf Teilkompetenzen im Zeitverlauf von der 6. bis zur 7. Klassenstufe betrachtet:

F3: Verändert sich die Kompetenzstruktur von Klassenstufe 6 bis 7, und wenn ja, in welcher Weise?

Mit dieser Frage soll untersucht und aufgezeigt werden, ob sich die Struktur der Rechtschreibkompetenz mit der längsschnittlichen Entwicklung der Rechtschreibkompetenz auf der Ebene der Personen- und Modellparametern ändert. Dazu werden

- die Entwicklung der Rechtschreibkompetenz auf der Ebene der fünf Teilkompetenzen modelliert, um die Kompetenzstruktur anhand der geschätzten Personenfähigkeiten zu beschreiben.
- tiefergehende Differenzierungen der Kompetenzstruktur durch ausgewählter Einflussfaktoren vorgenommen.
- Veränderungen der Kompetenzstruktur mittels von Korrelations- und Reliabilitätsanalysen und Modellgeltungstests festgestellt.

8.3.1 DATENGRUNDLAGE DER ENTWICKLUNGSSTUDIEN K6 UND K7

Aufbauend auf den Ergebnissen von Kapitel 8.2 erfolgt nun die längsschnittliche Modellierung des differenziellen Rechtschreibkompetenzmodells mit den Daten der Entwicklungsstudien (ES) der Klassenstufe 6 (K6) und 7 (K7). Analog zur längsschnittlichen Kompetenzmodellierung für das ganze Wort erfolgt eine getrennte Skalierung mit latenten Dimensionen. Infolgedessen werden die Entwicklungen der Schülerinnen und Schüler in den fünf Teilkompetenzen ermittelt, um eine differenzierte Aussage über die Beherrschung der Rechtschreibkompetenz zu ermöglichen. Dies wird zusätzlich erreicht, indem die ermittelten Kompetenzzuwächse nach ausgewählten Einflussfaktoren differenziert werden. Abschließend werden zur Bestimmung der Dimensionalität des Kompetenzkonstrukts Ergebnisse zur Korrelations- und Reliabilitätsstruktur und von unterschiedlichen Modellgeltungstests aufgezeigt.

8.3.1.1 LÄNGSSCHNITTLICHE MODELLIERUNG DER KOMPETENZSTRUKTUR

Die längsschnittliche Skalierung der fünf Teilkompetenzen für die ES K6 und K7 umfasst insgesamt 225 Struktureinheiten, die im Einzelnen der Tabelle 8.26 zu entnehmen sind.

Tabelle 8.26: Anzahl der Struktureinheiten und Ankeritems – ES K6 & K7

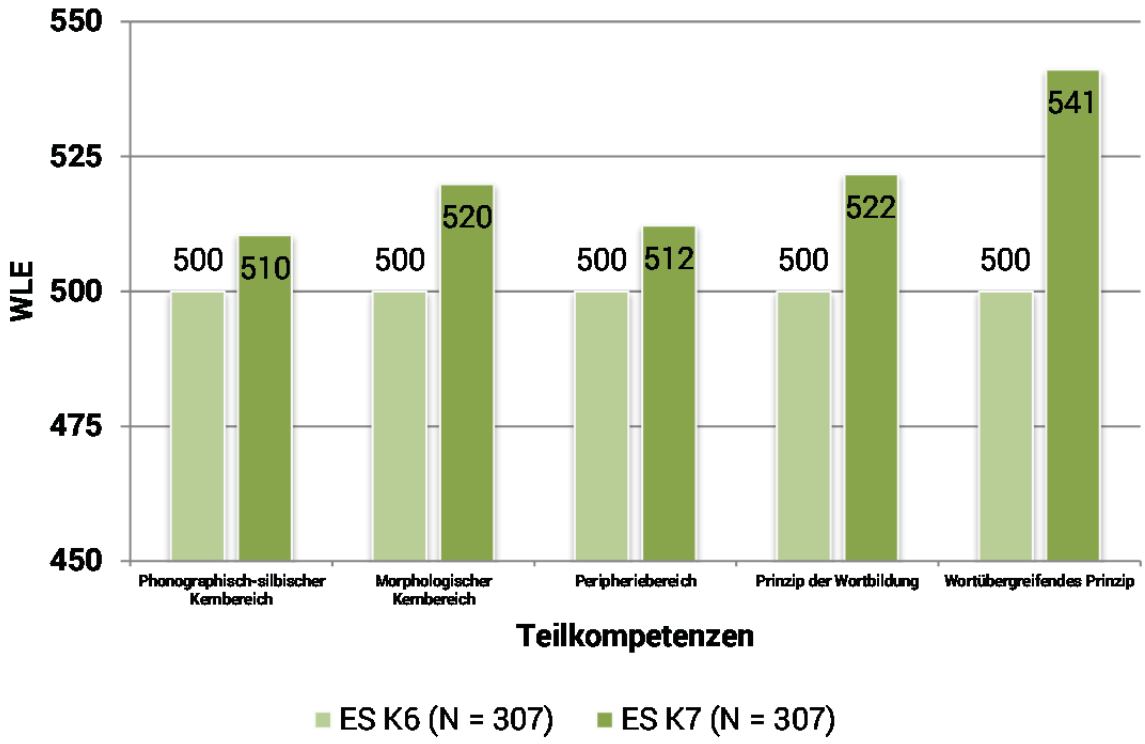
	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
Struktureinheiten	33	33	30	71[28]	58
Ankeritems aus ES K6	17	16	11	35	14
Ankeritems aus ES K7	16	17	19	36	44

Der Kernbereich, bestehend aus dem phonografisch-silbischen Prinzip und dem morphologischen Prinzip, umfasst jeweils 33 Ankeritems, wovon 17 bzw. 16 Struktureinheiten als Verankerung in der ES K6 dienen und 16 bzw. 17 weitere Einheiten bei der ES K7 hinzukommen. Die 30 Struktureinheiten zur Verankerung des Peripheriebereichs teilen sich in 11 und 19 Ankeritems für die ES K6 und K7 auf. Das Prinzip der Wortbildung umfasst insgesamt 71 Ankeritems, wovon sich 35 verankernde Einheiten aus der K6 und 36 Einheiten aus der K7 identifizieren lassen. Das wortübergreifende Prinzip beinhaltet 58 Ankeritems, die sich aus 14 Einheiten der K6 und weiteren 44 Einheiten der K7 ergeben.

Auf der Grundlage der 225 zur Verfügung stehenden Ankeritems wurde die längsschnittliche Modellierung des differenziellen Rechtschreibkompetenzmodells vorgenommen, wobei folgende Entwicklungen der fünf Teilkompetenzen bei den Schülerinnen und Schülern ermittelt werden konnten.

[28] Eine Struktureinheit wurde von allen Schülerinnen und Schülern richtig gelöst und wurde direkt ausgeschlossen.

Abbildung 8.31: Differenzielle Kompetenzentwicklung – ES K6 & K7



Die Schülerinnen und Schüler können sich sowohl im phonographisch-silbischen Prinzip als auch im morphologischen Prinzip verbessern. Dabei beträgt die Entwicklung im Fall des phonographisch-silbischen Prinzips 10 WLE-Punkte (d = -0.10) und im Fall des morphologischen Prinzips 20 WLE-Punkte (d = -0.19). Im Peripheriebereich konnte eine Verbesserung von 12 WLE-Punkten (d = -0.12) festgestellt werden. Das Prinzip der Wortbildung weist eine Verbesserung in Höhe von 22 WLE-Punkten (d = -0.20) auf. Weitaus höher fällt die Entwicklung beim wortübergreifenden Prinzip aus, da sich hier die Schülerinnen und Schüler mit 41 WLE-Punkten (d = -0.36) am stärksten verbessern. Diese Entwicklungen lassen sich in Anbetracht der Effektstärken bewerten, wonach beim wortübergreifenden Prinzip ein mittlerer Effekt vorliegt und ansonsten nur kleine Effekte beobachtet werden können.

Um einen tiefergehenden Einblick in die Entwicklung der fünf Teilkompetenzen zu erhalten, wird nachfolgend eine Differenzierung unter Zuhilfenahme ausgewählter Einflussfaktoren vorgestellt.

8.3.1.2 DIFFERENZIERUNG DER DIFFERENZIELLEN KOMPETENZENTWICKLUNG

Zu den berücksichtigten Einflussfaktoren zählen das Geschlecht, das Alter, der Migrationshintergrund und die Verkehrssprache, der sozioökonomische Status sowie die jeweils besuchte Schulform. Mit ihrer Hilfe wird die längsschnittliche Entwicklung der Kompetenzstruktur für die Entwicklungsstudien (ES) von der Klassenstufe 6 bis 7 (K6/K7) vertiefend ausdifferenziert.

Zunächst wird die differenzielle Kompetenzentwicklung hinsichtlich des Geschlechts der Schülerinnen und Schüler dargestellt. Die 164 Jungen und 142 Mädchen erreichen die in Tabelle 8.27 dargestellten Kompetenzwerte bei den fünf Teilkompetenzen.

Tabelle 8.27: Differenzielle Kompetenzentwicklung nach Geschlecht – ES K6 & K7

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
			WLE-Punkte		
Jungen (N = 164)	492[A]	501[A]	494[A]	498[A]	516[A]
Mädchen (N = 142)	532[A]	542[A]	534[A]	550[A]	571[A]
Signifikante Unterschiede mit [A] gekennzeichnet					

Zweifelsfrei ist zu erkennen, dass sich die Kompetenzentwicklung bei Jungen und Mädchen unterscheidet. Die insgesamt signifikanten Unterschiede zwischen den Geschlechtern verteilen sich im Kernbereich folgendermaßen: Während beim phonographisch-silbischen Prinzip die Mädchen 32 WLE-Punkte zulegen, verringert sich die Kompetenz der Jungen um acht WLE-Punkte. Dies zeigt sich auch bei dem morphologischen Prinzip, wo es zu einer Entwicklung in Höhe von einem WLE-Punkt bzw. 42 WLE-Punkten bei den Jungen und Mädchen kommt. Beim Peripheriebereich verringert sich die Kompetenz der Jungen um acht WLE-Punkte und bei den Mädchen steigt die Kompetenz um 34 WLE-Punkte an. Noch deutlicher fallen die Unterschiede zwischen den Geschlechtern bei dem Prinzip der Wortbildung und dem wortübergreifenden Prinzip auf. Die Jungen weisen eine Verringerung der Kompetenz um zwei WLE-Punkte bei dem Prinzip der Wortbildung auf und erzielen bei dem wortübergreifenden Prinzip die höchste Entwicklung mit 16 WLE-Punkten. Die Mädchen hingegen zeigen auch an dieser Stelle eine deutlich positive Entwicklung, die beim Prinzip der Wortbildung und dem

wortübergreifenden Prinzip mit 50 bzw. 71 WLE-Punkten mit Abstand am höchsten ausfällt. Demnach kann für den Längsschnitt der ES K6 und K7 festgehalten werden, dass sich die Mädchen in ihrer Rechtschreibkompetenz positiv entwickeln, während sich bei den Jungen eine Stagnation oder sogar rückläufige Tendenzen zeigen.

Als weiteres wichtiges Differenzierungskriterium wird das Alter der Schülerinnen und Schüler[29] berücksichtigt, um altersspezifische Unterschiede bei der Entwicklung der Teilkompetenzen herauszustellen.

Tabelle 8.28: Differenzielle Kompetenzentwicklung nach Alter – ES K6 & K7

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
			WLE-Punkte		
14 Jahre (N = 99)	542^{AC}	546^{A}	532^{A}	546^{AC}	566^{A}
15 Jahre (N = 159)	510^{BC}	518	512	526^{BC}	545^{B}
16 Jahre (N = 19)	432^{ABC}	458^{A}	463^{A}	457^{ABC}	458^{AB}
Signifikante Unterschiede nach Scheffé mit ^{ABC} gekennzeichnet					

In Bezug auf das Alter zeigen sich insbesondere bei den 14 und 15 Jahre alten Schülerinnen und Schülern Kompetenzzuwächse, wobei die Kompetenz bei den 16-Jährigen abnimmt. Die Schülerinnen und Schüler im Alter von 14 Jahren weisen eine Entwicklung zwischen 32 und 66 WLE-Punkten bei den Teilkompetenzen auf, die am geringsten im Peripheriebereich und am stärksten beim wortübergreifenden Prinzip ausfällt. Geringer ist der Kompetenzzuwachs bei den 15-Jährigen, die sich lediglich zwischen 10 und 45 WLE-Punkten in den Teilkompetenzen verbessern. Es fällt auf, dass bei ihnen im Kern- und Peripheriebereich die geringste Entwicklung stattfindet. Die rückläufigen Entwicklungen der Teilkompetenzen bei den Schülerinnen und Schülern im Alter von 16 Jahren betragen zwischen 37 und 68 WLE-Punkten. Mit zunehmendem Alter der Schülerinnen und Schüler nimmt die Intensität der Kompetenzentwicklung

[29] Fünf Angaben werden von den Analysen ausgeschlossen, da sie pro Kategorie eine zu geringe Fallzahl aufweisen.

im Bereich der Rechtschreibung ab. Signifikante Unterschiede ergeben sich besonders beim phonographisch-silbischen Prinzip und beim Prinzip der Wortbildung.

Der Migrationshintergrund wird als möglicher Einflussfaktor anhand des Geburtslandes der Eltern und der im Haushalt gesprochenen Sprache untersucht und im Vergleich mit der Kompetenzentwicklung von Schülerinnen und Schülern ohne Migrationshintergrund dargestellt.

Tabelle 8.29: Differenzielle Kompetenzentwicklung nach Migrationshintergrund – ES K6 & K7

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
			WLE-Punkte		
beide Elternteile (N = 52)	502	512	519	522	533
ein Elternteil (N = 35)	478	513	483	494	521
kein Elternteil (N = 212)	516	521	515	524	547

Die Kompetenzentwicklung unterscheidet sich hinsichtlich des Migrationshintergrundes der Schülerinnen und Schüler, bezogen auf das Geburtsland der Eltern, jedoch nicht signifikant. Die Kompetenz entwickelt sich höher, wenn die Eltern in Deutschland oder beide im Ausland geboren sind. Ist dagegen nur ein Elternteil im Ausland geboren, sind bei diesen Schülerinnen und Schülern nur geringe Zuwächse bzw. rückläufige Tendenzen zu verzeichnen.

Tabelle 8.30: Differenzielle Kompetenzentwicklung nach Sprache im Haushalt – ES K6 & K7

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
			WLE-Punkte		
Deutsch (N = 271)	512	522	514	524	545
andere Sprache (N = 24)	497	509	489	500	521

Ebenfalls zeigen sich im Hinblick auf die Sprache im Haushalt Auswirkungen auf die Kompetenzentwicklung, die allerdings nicht signifikant sind. Tendenziell fällt auf, dass die Kompetenzentwicklung bei Schülerinnen und Schülern, die Deutsch zu Hause sprechen, höher ausfällt als bei denjenigen mit einer anderen Sprache im Haushalt.

Inwiefern sich der sozioökonomische Status auf die Entwicklung der Teilkompetenzen niederschlägt, wird anhand der im Haushalt zur Verfügung stehenden Bücher aufgezeigt.

Tabelle 8.31: Differenzielle Kompetenzentwicklung nach Anzahl der Bücher – ES K6 & K7

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
			WLE-Punkte		
keine oder nur sehr wenige (0-10 Bücher) (N = 13)	433^{A}	487	443^{A}	440^{A}	443^{A}
genug, um ein Regalbrett zu füllen (11-25 Bücher) (N = 41)	480^{B}	504	493	508	494^{B}
genug, um mehrere Regalbretter zu füllen (26-100 Bücher) (N = 84)	494^{C}	504	496^{B}	505^{B}	513^{C}
genug, um ein kleines Regal zu füllen (101-200 Bücher) (N = 67)	518	526	514	515	538^{D}
genug, um ein großes Regal zu füllen (201-500 Bücher) (N = 57)	560^{ABC}	551	551^{AB}	569^{AB}	616^{ABCD}
genug, um eine Regalwand zu füllen (mehr als 500 Bücher) (N = 35)	541	543	550^{A}	560	590^{AB}
Signifikante Unterschiede nach Scheffé mit ^{ABCD} gekennzeichnet					

Gestaffelt nach der Anzahl der Bücher im Haushalt ist zu erkennen, dass sich die Kompetenzentwicklung verändert. Sind keine oder nur sehr wenige Bücher vorhanden, führt dies zu einer deutlich negativen Entwicklung der Teilkompetenzen. Bei bis zu 100 Büchern im Haushalt sind geringe Zuwächse bei der Kompetenzentwicklung zu verzeichnen. Erst im Bereich von über 100 bis 500 Büchern zeigen sich deutliche Kompetenzzuwächse. Bei mehr als 500 Büchern fällt die Entwicklung in den Teilkompetenzen dagegen geringer aus. Mit Ausnahme des morphologischen Prinzips zeigen sich signifikante Unterschiede bei der Entwicklung der Teilkompetenzen. Besonders deutliche Unterschiede bestehen beim wortübergreifenden Prinzip. Ins-

besondere die Gruppen bis zu 100 Bücher im Haushalt unterscheiden sich signifikant von der Gruppe mit 201 bis 500 Büchern. So lässt sich folgern, dass die im Haushalt zur Verfügung stehende Anzahl an Büchern Auswirkungen auf die Kompetenzentwicklung hat. Diese sind vermutlich auf die Bildungsorientierung der Familie zurückzuführen (vgl. Bos, Schwippert & Stubbe, 2007).

Abschließend wird betrachtet, inwieweit sich durch den Besuch unterschiedlicher Schulformen Veränderungen in der Kompetenzentwicklung ergeben.

Tabelle 8.32:Differenzielle Kompetenzentwicklung nach Schulform – ES K6 & K7

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
			WLE-Punkte		
Hauptschule (N = 40)	423^{AE}	440^{AD}	428^{ACE}	408^{ABCE}	429^{ACE}
SMB (N = 53)	448^{BE}	478^{BD}	457^{BDE}	472^{ABE}	454^{BE}
Realschule (N = 22)	492^{CDE}	518^{ACD}	503^{ACDE}	501^{ACDE}	512^{ACDE}
Gesamtschule (N = 19)	395^{CDE}	397^{BCD}	404^{CDE}	413^{CDE}	404^{CDE}
Gymnasium (N = 173)	565^{ABCDE}	565^{ABD}	562^{ABCDE}	578^{ABCDE}	612^{ABCDE}
Signifikante Unterschiede nach Scheffé mit ^{ABCDE} gekennzeichnet					

Die Schulform stellt den Einflussfaktor dar, der am stärksten auf die differenzielle Kompetenzentwicklung einwirkt. Die Unterscheide bei den Teilkompetenzen hinsichtlich der Schulformen sind überwiegend signifikant. Demnach zeigt sich deutlich, dass sich die Schülerinnen und Schüler in den jeweiligen Schulformen unterschiedlich entwickeln. So ergibt sich bei den Gesamtschülerinnen und -schülern eine deutliche Verringerung der Teilkompetenzen, wohingegen die Gymnasiastinnen und Gymnasiasten ihre Kompetenz deutlich erhöhen können. Zwischen diesen Extremen ordnen sich aufsteigend die Schülerinnen und Schüler an Hauptschulen, an Schulen mit mehreren Bildungsgängen (SMB) und an Realschulen ein.

Zusammenfassend konnte nachgewiesen werden, dass die untersuchten Einflussfaktoren die differenzielle Kompetenzentwicklung in unterschiedlichem Maße beeinflussen.

Nach der Ausdifferenzierung der differenziellen Kompetenzentwicklung im Hinblick auf Einflussfaktoren wird weitergehend aufgezeigt, wie sich die Dimensionalität der Kompetenzstruktur zu ihrer längsschnittlichen Entwicklung verhält.

8.3.1.3 DIMENSIONALITÄT DES KOMPETENZKONSTRUKTS

Anhand der Modellparameter wird abschließend der Zusammenhang zwischen der Entwicklung der Rechtschreibkompetenz bei den Schülerinnen und Schülern von der 6. bis zur 7. Klassenstufe und deren Auswirkungen auf die Kompetenzstruktur dargestellt. Dazu wird auf der Grundlage der Korrelations- und Reliabilitätsstruktur und der Modellgeltungstests überprüft, ob sich im Längsschnitt Veränderungen der Kompetenzstruktur ergeben.

Zunächst erfolgt die Darstellung der Korrelations- und Reliabilitätsanalysen in einer gemeinsamen Betrachtung der beiden Messzeitpunkte in Klassenstufe 6 und 7, um so Veränderungen an der Kompetenzstruktur feststellen zu können.

Abbildung 8.32: Latente Korrelationen und Reliabilitäten der fünf Teilkompetenzen – ES K6 & K7

	Phonographisch-silbischer Kernbereich	Morphologischer Kernbereich	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
Phonographisch-silbischer Kernbereich	0.93 0.90	0.99	0.98	0.99	0.88
Morphologischer Kernbereich	0.98	0.93 0.89	0.99	0.99	0.85
Peripheriebereich	0.96	0.97	0.91 0.91	0.96	0.83
Prinzip der Wortbildung	0.96	0.95	0.95	0.93 0.91	0.88
Wortübergreifendes Prinzip	0.83	0.82	0.85	0.87	0.90 0.89

Latente Interkorrelationen (dargestellt auf der oberen und unteren Dreiecksmatrix) sowie Reliabilitäten (EAP/PV) (dargestellt auf der Diagonalen) der fünf Teilfähigkeiten für den Längsschnitt (N = 307) der Entwicklungsstudie K6 (obere Angaben) und der Entwicklungsstudie K7 (untere Angaben).

Sowohl in der ES K6 als auch in der K7 korrelieren die fünf Teilkompetenzen stark miteinander, das sich an den hohen Korrelationskoeffizienten zwischen 0.83 und 0.99 erkennen lässt. Während sich vergleichbar hohe Zusammenhänge der ersten vier Teilkompetenzen zeigen, sind die Korrelationen zum wortübergreifenden Prinzip geringer und variieren im Bereich von 0.82 und 0.88. Im Gegensatz zur ES K6 sind die Zusammenhänge bei der ES K7 geringer, wobei die leicht gestiegene Korrelation zwischen dem Peripheriebereich und dem wortübergreifenden Prinzip eine Ausnahme darstellt. Diese Entwicklung ergibt sich auch bei den Reliabilitäten, die bis zu einem Wert von 0.04 bei der ES K7 geringer sind. Analog zu den in Kapitel 8.1 vorgestellten Ergebnissen können auch an dieser Stelle trotz hoher Zusammenhänge zwischen den jeweiligen Teilkompetenzen zufriedenstellende Reliabilitäten nachgewiesen werden.

Aus diesem Grund sind nachfolgend die Modellgeltungstests in Tabelle 8.33 von besonderer Bedeutung, um in diesem Zusammenhang die empirische Modellierung kritisch zu reflektieren und herauszufinden, ob sich im Längsschnitt ebenfalls Veränderungen beim Modellieren der Kompetenzstruktur ergeben.

Tabelle 8.33: Modellgeltungstests der mehrdimensionalen Skalierung – ES K6 & K7

	1D	2D	4D	5D
Deviance	39639.04401 39617.52782	39474.62389 39441.60985	39434.86085 39394.31831	**39430.89343** **39339.32852**
AIC	40091.04401 40069.52782	39930.62389 39897.60985	**39904.86085** 39864.31831	39910.89343 **39819.32852**
CAIC	41159.31160 41137.79541	**41008.34518** 40975.33114	41015.67007 **40975.12753**	41045.33689 40953.77198
BIC	40933.31160 40911.79541	**40780.34518** 40747.33114	40780.67007 40740.12753	40805.33689 **40713.77198**

ES K6 (N = 307, 225 Struktureinheiten)
ES K7 (N = 307, 225 Struktureinheiten)

Wie bereits in Kapitel 5.2 und 8.1.3 aufgezeigt wurde, weist der Modellgeltungstest auf der Grundlage der Deviance-Statistik das fünfdimensionale Modell für die beste Präsentation der in den Daten enthaltenen Informationen aus. Diese Eindeutigkeit geht im Fall des Längsschnitts der ES K6 und K7 verloren, wenn neben der Deviance-Statistik noch die Anzahl der Parameter und die Stichprobengröße bei den weiteren Modellgeltungstests AIC, CAIC und BIC mit einbezogen werden. Demnach ist ansatzweise zu erkennen, dass die Datenstruktur mit weniger Dimensionen zu modellieren ist. So ist für die ES K6 tendenziell eher eine zwei- oder vierdimensionale Modellierung des differenziellen Kompetenzkonstrukts abzuleiten. Allerdings zeigt sich dies nicht für die ES K7, wo erneut die fünfdimensionale Datenstruktur dominiert und sich nur bei der CAIC-Statistik ein Hinweis für eine geringere Dimensionalität ergibt. Aus diesem Grund ist die Befundlage weiterhin nicht eindeutig und es bleibt offen, wie die Dimensionalität des Kompetenzkonstrukts idealerweise längsschnittlich empirisch modelliert werden muss. Weiterführende Einsichten sind von längsschnittlichen Analysen der Haupterhebungsdaten aus den Klassenstufen 5, 7 und 9 zu erwarten.

8.4 ERGEBNISSE ZUR ERMITTLUNG VON EINFLUSSFAKTOREN

Im Weiteren werden mögliche Einflüsse auf die Kompetenzentwicklung beim ganzen Wort untersucht. Dazu werden zum einen die Entwicklung in den Teilkompetenzen und zum anderen die in dieser Arbeit berücksichtigten Einflussfaktoren herangezogen. Dies ist inhaltlich von großem Interesse, da die gemessene Personenfähigkeit beim ganzen Wort zum Teil weitaus geringer ist als in den Teilkompetenzen.

Für die Untersuchung der vierten Forschungsfrage

F4: Beeinflussen die Entwicklung in den Teilkompetenzen bzw. die Faktoren Geschlecht, Alter, sozioökonomischer Hintergrund, Migrationshintergrund und Schulform die Entwicklung der Rechtschreibkompetenz auf der Ebene des ganzen Wortes von Klassenstufe 6 und 7, und wenn ja, in welcher Weise?

Werden Korrelations- und Regressionsanalysen verwendet, um

– zunächst zu klären, in welchem korrelativen Zusammenhang die ermittelte Kompetenzentwicklung beim ganzen Wort mit den fünf Teilkompetenzen und den Einflussfaktoren steht.

– die Kompetenzwicklung des ganzen Wortes regressionsanalytisch durch die Teilkompetenzen bzw. Einflussfaktoren zu erklären.

8.4.1 KORRELATIONS- UND REGRESSIONSANALYSE

Für die Durchführung der Analysen wird anhand der Ankeritems die gemessene Differenz der Personenfähigkeit zwischen den beiden Messzeitpunkten auf der Ebene des ganzen Wortes als abhängige bzw. zu erklärende Variable (AV) herangezogen, um sie durch die unabhängigen Variablen (UV) in Form der Kompetenzwertdifferenzen für die fünf Teilkompetenzen und der sechs Einflussfaktoren mithilfe von Korrelations- und Regressionsanalysen als prädiktive Faktoren zu identifizieren. In einem ersten Schritt wird dazu getrennt voneinander korrelations- und regressionsanalytisch geprüft, inwiefern die Teilkompetenzen bzw. die Einflussfaktoren zur Aufklärung der Kompetenzentwicklung beim ganzen Wort beitragen. Die abschließende Regressionsanalyse berücksichtigt gemeinsam die Teilkompetenzen des differenziellen Kompetenzkonstrukts und die Einflussfaktoren.

8.4.1.1 PRÜFUNG DES ZUSAMMENHANGS DER TEILKOMPETENZEN MIT DER KOMPETENZENTWICKLUNG

Im Gegensatz zu den latenten Korrelationen des differenziellen Kompetenzkonstrukts aus Kapitel 8.3 fallen die manifesten Korrelationskoeffizienten in Tabelle 8.34 der Produkt-Moment-Korrelation nach Pearson (Kapitel 3.4) geringer aus, da sie lediglich die manifesten Zusammenhänge zwischen der Kompetenzentwicklung auf der Ebene des ganzen Wortes und den fünf Teilkompetenzen abbilden.

Tabelle 8.34: Manifeste Korrelationen der Teilkompetenzen – ES K6 & K7

	Ganzes Wort	Phonographisch-silbisches Prinzip	Morphologisches Prinzip	Peripheriebereich	Prinzip der Wortbildung	Wortübergreifendes Prinzip
Ganzes Wort	1	.793**	.785**	.910**	.885**	.863**
Phonographisch-silbisches Prinzip	.793**	1	.718**	.748**	.733**	.661**
Morphologisches Prinzip	.785**	.718**	1	.738**	.727**	.648**
Peripheriebereich	.910**	.748**	.738**	1	.798**	.747**
Prinzip der Wortbildung	.885**	.733**	.727**	.798**	1	.728**
Wortübergreifendes Prinzip	.863**	.661**	.648**	.747**	.728**	1

N= 307
** Pearson-Korrelation ist bei Niveau 0.01 signifikant (zweiseitig).
* Pearson-Korrelation ist bei Niveau 0.05 signifikant (zweiseitig).

Trotz der im Verhältnis geringeren manifesten Korrelationen bestehen zwischen dem ganzen Wort und den Teilkompetenzen starke bis sehr starke Zusammenhänge, deren Stärke im Bereich von 0.64 bis 0.91 variiert. Es zeigt sich, dass die Entwicklung der Personenfähigkeit der Schülerinnen und Schüler in Bezug auf das ganze Wort weniger mit dem Kernbereich, bestehend aus dem phonographisch-silbischen (.793) und dem morphologischen Prinzip (.785), zusammenhängt, sondern insbesondere mit dem Peripheriebereich (.910) und dem Prinzip der Wortbildung (.885) sowie dem wortübergreifenden Prinzip (.863) höhere Korrelationen bestehen. Diese Befundlage lässt für den Längsschnitt von der Klassenstufe 6 bis 7 vermuten, dass in den höheren Klassenstufen der Zuwachs im Kernbereich geringer ausfällt zugunsten einer stärkeren Ausprägung der erweiterten Teilkompetenzen. Ebenso korrelieren die Teilkompetenzen stark untereinander, weshalb von einem zusammenhängenden Kompetenzkonstrukt ausgegangen werden kann, das sich in unterschiedlichem Maße an der Kompetenzentwicklung insgesamt beteiligt.

Inwiefern die Entwicklung der Rechtschreibkompetenz auf der Ebene des ganzen Wortes anhand der fünf Teilkompetenzen erklärt wird, zeigt die dazugehörige schrittweise Regressionsanalyse in Tabelle 8.35.

Tabelle 8.35: Regressionsanalyse der Teilkompetenzen – ES K6 & K7

	Modell 1 B (SE) Beta	**Modell 2** B (SE) Beta	**Modell 3** B (SE) Beta	**Modell 4** B (SE) Beta	**Modell 5** B (SE) Beta
Konstante	**10.93 (3.68)*** **	5.17 (3.22)	**4.93 (2.22)** **	**2.20 (1.88)***	**-3.82 (1.55)** **
Phonographisch-silbisches Prinzip	**0.73 (.03) 0.79*** **	**0.43 (.04) 0.47*** **	**0.17 (.03) 0.19*** **	**0.10 (.03) 0.11*** **	**0.07 (.02) 0.07*** **
Morphologisches Prinzip		**0.44 (.04) 0.45*** **	**0.18 (.03) 0.18*** **	**0.10 (.03) 0.10*** **	**0.08 (.02) 0.08*** **
Peripheriebereich			**0.68 (.04) 0.64*** **	**0.50 (.03) 0.47*** **	**0.38 (.03) 0.36*** **
Prinzip der Wortbildung				**0.33 (.03) 0.36*** **	**0.25 (.02) 0.27*** **
Wortübergreifendes Prinzip					**0.24 (.02) 0.30*** **
R^2	**0.628**	**0.724**	**0.870**	**0.908**	**0.943**

N = 307
ns = nicht signifikant; * = signifikant (p<.05); ** = signifikant (p<.01) ; *** = signifikant (p<.001)

Je nach Kompetenzausprägung beim phonographisch-silbischen Prinzip verändert sich beim ersten Modell die Entwicklung der Personenfähigkeit der Schülerinnen und Schüler beim ganzen Wort mit einem Effekt von 0.73. Mit knapp 63 Prozent wird die Entwicklung der Rechtschreibkompetenz auf der Ebene des ganzen Wortes bereits mit dem phonographisch-silbischen Prinzip erklärt. Wird das morphologische Prinzip beim zweiten Modell mit aufgenommen, verringert sich der vorherige Effekt (0.43) und die Entwicklung beim ganzen Wort wird durch das morphologische Prinzip zu 0.44 beeinflusst. Somit klärt der Kernbereich, bestehend aus dem phonographisch-silbischen und morphologischen Prinzip, die Entwicklung der Kompetenz beim ganzen Wort zu ungefähr 72 Prozent auf. Durch die Hinzunahme des Peripheriebereichs wird ein Effekt in Höhe von 0.68 beim dritten Modell nachgewiesen, wodurch die Effekte des Kernbereichs deutlich kleiner werden. Unter Berücksichtigung des Peripheriebereichs verliert der Kernbereich an Bedeutung für die Entwicklung der Personenfähigkeit beim ganzen Wort, die durch den Kern- und Peripheriebereich zu 87 Prozent erklärt werden kann. Das Prinzip der Wortbildung im vierten Modell hat einen Effekt von 0.33, der zur Entwicklung der Rechtschreibkompetenz beim ganzen Wort beiträgt, wobei sich die weiteren enthaltenden Teilkompetenzen in ihren Effekten verringern. Mit knapp 91 Prozent erklären die für die letztgenannte Analyse im Modell aufgenommenen vier unabhängigen Variablen die

abhängige Variable. Das fünfte Modell beinhaltet alle fünf Teilkompetenzen als erklärende Variablen, die die Entwicklung der Kompetenz des ganzen Wortes als abhängige Variable mit knapp 94 Prozent erklären. Dabei weist das wortübergreifende Prinzip als fünfte erklärende Variable einen Effekt in Höhe von 0.24 auf, wobei sich die Effekte der übrigen vier unabhängigen Variablen weiter verringern. Somit besitzt der Kernbereich die geringsten Effekte in Höhe von maximal 0.08, das Prinzip der Wortbildung und das wortübergreifende Prinzip haben einen ungefähren Effekt von 0.25 und der Peripheriebereich zeigt mit 0.38 den größten Effekt.

Zusammenfassend ergibt sich aufgrund der Ergebnisse zur Regressionsanalyse das Muster, das bereits bei der Korrelationsanalyse zum Vorschein kam, d. h., dass die abhängige Variable bei Betrachtung aller Teilkompetenzen vor allem durch die Teilkompetenz im Peripheriebereich und in der Wortbildung sowie im wortübergreifenden Prinzip und nicht durch die Teilkompetenz im Kernbereich beeinflusst werden. Diese Analyseergebnisse sind von einem hohen didaktischen Interesse. Sie legen den Schluss nahe, dass die Kompetenzentwicklung im Hinblick auf eine korrekte Rechtschreibung in der Sekundarstufe I in erster Linie von Einsichten in die morphologische und syntaktische Struktur der Schrift sowie von automatisierten Schreibungen im Peripheriebereich bestimmt wird. Dabei handelt es sich vorwiegend um Lerninhalte des Rechtschreibunterrichts in der Sekundarstufe I. Eine Beherrschung des Kernbereichs alleine, der vorwiegend Lerngegenstand in der Grundschule ist, führt dagegen nicht weiter.

8.4.1.2 PRÜFUNG DES ZUSAMMENHANGS DER EINFLUSSFAKTOREN MIT DER KOMPETENZENTWICKLUNG

Welche Faktoren weitere Effekte auf die Entwicklung der Kompetenz beim ganzen Wort ausüben, wird nun korrelations- und regressionsanalytisch für die genannten Einflussfaktoren untersucht und dargestellt.

Tabelle 8.36: Manifeste Korrelationen der Einflussfaktoren – ES K6 & K7

	Ganzes Wort	Geschlecht	Alter	Migrationshintergrund	Sprache im Haushalt	Anzahl der Bücher	Schulform
Ganzes Wort	1	**-.217****	**-.201****	-.057	.079	**.307****	**.612****
Geschlecht	**-.217****	1	.076	.002	-.005	.014	-.020
Alter	**-.201****	.076	1	-.023	-.003	**-.146***	**-.150***
Migrationshintergrund	-.057	.002	-.023	1	**-.476****	**-.238****	-.022
Sprache im Haushalt	.079	-.005	-.003	**-.476****	1	.114	.035
Anzahl der Bücher	**.307****	.014	**-.146***	**-.238****	.114	1	**.301****
Schulform	**.612****	-.020	**-.150***	-.022	.035	**.301****	1

N = 267 (min.) - 307 (max.)
** Pearson-Korrelation ist bei Niveau 0.01 signifikant (zweiseitig).
* Pearson-Korrelation ist bei Niveau 0.05 signifikant (zweiseitig).

Im Hinblick auf die untersuchten Einflussfaktoren zeigen sich Zusammenhänge, die sich hinderlich bzw. förderlich auf die Entwicklung der Rechtschreibkompetenz auf der Ebene des ganzen Wortes auswirken. Als hinderliche Faktoren wurden das Geschlecht (-.217) und Alter (-.201) identifiziert, die einen signifikant schwachen negativen Zusammenhang aufweisen. Förderlich wirken sich dagegen die im Haushalt zur Verfügung stehenden Bücher (.307) und die jeweilige Schulform (.612) aus, die einen schwachen bis mittleren positiven Zusammenhang mit der Kompetenzentwicklung aufweisen. Für die Entwicklung der Rechtschreibkompetenz auf der Ebene des ganzen Wortes sind der Migrationshintergrund und die im Haushalt gesprochene Sprache ohne signifikante Bedeutung.

Die Einflussfaktoren korrelieren untereinander nur wenig. Abhängig vom Alter unterscheidet sich der Einfluss, den die Anzahl der Bücher im Haushalt (-.146) und der Besuch der jeweiligen Schulformen (-.150) nehmen. Sofern ein Migrationshintergrund vorliegt, werden der Gebrauch

einer anderen Sprache als Deutsch (-.476) begünstigt und die Anzahl im Haushalt zur Verfügung stehenden Bücher (-.238) verringert. Bei der Schulform liegt ein eindeutiger Zusammenhang mit der Anzahl der Bücher im Haushalt vor.

Die Ergebnisse der Regressionsanalyse zu den untersuchten Einflussfaktoren werden weitergehend aufgezeigt.

Tabelle 8.37: Regressionsanalyse der Einflussfaktoren – ES K6 & K7

	Modell 1 B (SE) Beta	**Modell 2** B (SE) Beta	**Modell 3** B (SE) Beta	**Modell 4** B (SE) Beta	**Modell 5** B (SE) Beta	**Modell 6** B (SE) Beta
Konstante	**46.85(9.31)*** **	**510.40 (161.21)*** **	**515.65 (161.35)*** **	**474.41 (162.48)*** **	**378.84 (165.23)** **	128.62 (127.45)
Geschlecht[A]	**-38.79 (12.85) -0.19*** **	**-34.84 (12.74) -0.17** **	**-34.77 (12.74) -0.17** **	**-35.71 (12.74) -0.17** **	**-36.87 (12.62) -0.18*** **	**-40.29 (9.63) -0.19*** **
Alter[B]		**-31.67 (11.00) -0.18*** **	**-31.79 (11.00) -0.18*** **	**-31.63 (10.98) -0.18*** **	**-26.97 (11.02) -0.15** **	-12.50 (8.48) -0.07
Migrationshintergrund[C]			13.21 (14.31) -0.06	2.32 (16.33) -0.01	1.70 (16.16) 0.01	-11.17 (12.37) -0.05
Sprache im Haushalt[D]				37.00 (26.88) 0.10	29.39 (26.77) 0.08	23.74 (20.43) 0.06
Anzahl der Bücher[E]					**45.52 (16.94) 0.15** **	8.50 (13.17) 0.03
Schulform[F]						**135.86 (10.14) 0.64*** **
R^2	**0.035**	**0.066**	**0.069**	**0.076**	**0.099**	**0.477**

N = 255
B (SE) Beta = unstandardisiertes Regressionsgewicht (Standardfehler) standardisiertes Regressionsgewicht
ns = nicht signifikant; * = signifikant (p<.05); ** = signifikant (p<.01) ; *** = signifikant (p<.001)
A = Geschlecht der Schülerinnen und Schüler (0 = weiblich; 1 = männlich)
B = Alter der Schülerinnen und Schüler (offene Angabe)
C = Migrationshintergrund (0 = mindestens ein Elternteil; 1 = kein Elternteil)
D = Sprache im Haushalt (0 = andere Sprache; 1 = Deutsch)
E = Anzahl der Bücher (0 = 101 und mehr Bücher; 1= 100 oder weniger Bücher)
F = Schulform Gymnasium (0 = andere Schulform; 1 = Gymnasium)

Die Kompetenzentwicklung beim ganzen Wort wird durch das Geschlecht der Schülerinnen und Schüler beeinflusst, da ein Effekt in Höhe von -38.79 vorliegt, was bedeutet, dass die Entwicklung der Jungen niedriger ausfällt. Allerdings werden mit dem Geschlecht nur knapp 4 Prozent der Varianz erklärt. Wird zusätzlich das Alter kontrolliert, verringert sich die zu erwartende Kompetenzentwicklung nochmals. Mit einem Effekt von -31.67 und einer Steigerung der aufgeklärten Varianz um drei Prozent wirkt sich das Alter auf die Kompetenzentwicklung aus, d. h., dass sich ältere Schülerinnen und Schüler weniger gut entwickeln bzw. sogar rückläufige Entwicklungen aufweisen. Zur besseren Entwicklung kommt es, sofern kein Migrationshintergrund vorliegt. Ist kein Elternteil im Ausland geboren, wirkt sich dies mit einem verhältnismäßig geringen, nicht signifikanten Effekt von 13.21 aus, was die erklärte Varianz nur geringfügig ändert. Dagegen ist es tendenziell vorteilhaft, wenn die Familiensprache Deutsch ist, da hier ein positiver Effekt in Höhe von 37.00 vorliegt. Jedoch ist auch dieser Befund nicht signifikant

und trägt wenig zur erklärten Varianz bei. Das fünfte Modell wird durch die unabhängige Variable der im Haushalt zur Verfügung stehenden Bücher ergänzt, die einen signifikanten Effekt von 45.52 bei 101 und mehr Büchern aufweist. Dadurch wird der signifikante Effekt des Alters leicht reduziert. Die Anzahl der Bücher ist ein guter Prädiktor zur Aufklärung der Varianz, die durch das fünfte Modell zu ungefähr 10 Prozent erklärt werden kann. Erweitert wird das sechste Modell durch den letzten Einflussfaktor: die Schulform, die die Schülerinnen und Schüler besuchen. Hierbei zeigt sich, dass die Entwicklung der Rechtschreibkompetenz auf der Ebene des ganzen Wortes mit einem Effekt von 135.86 beeinflusst wird, wenn es sich bei der besuchten Schulform um ein Gymnasium handelt. Die Schulform hat zudem starke Auswirkungen auf die erklärte Varianz von nunmehr knapp 48 Prozent. Ebenso zeigen sich Veränderungen bei den weiteren Effekten der Einflussfaktoren. Die Kompetenzentwicklung ist demnach etwas stärker vom Geschlecht abhängig und weniger vom Alter, da der gesunkene Effekt nicht mehr signifikant ist. Die Effekte eines Migrationshintergrundes und der Gebrauch der deutschen Sprache im Haushalt verringern sich bei diesem Modell wie auch der Effekt, der aus der Anzahl der Bücher im Haushalt resultiert.

Insgesamt lässt sich anhand der Ergebnisse festhalten, dass die Kompetenzentwicklung beim ganzen Wort am stärksten vom Geschlecht der Schülerinnen und Schüler und dem Besuch der jeweiligen Schulform beeinflusst wird. Nicht prädiktiv sind dagegen das Alter, der Migrationshintergrund, die Sprache im Haushalt und der Besitz von Büchern, die keinen eindeutig nachweisbaren Effekt auf die Kompetenzentwicklung haben.

8.4.1.3 PRÜFUNG DER TEILKOMPETENZEN UND EINFLUSSFAKTOREN

Durch die Kombination der bislang getrennt durchgeführten Regressionsanalysen wird abschließend aufgezeigt, wie sich die Entwicklung der Schülerfähigkeit beim ganzen Wort durch die Kompetenzentwicklung in den fünf Teilkompetenzen unter Kontrolle der Einflussfaktoren erklären lässt. Dabei wird auf eine schrittweise Darstellung der Regressionsmodelle verzichtet, um ein Modell mit allen Teilkompetenzen und Einflussfaktoren zu präsentieren.

Tabelle 8.38: Regressionsanalyse der Teilkompetenzen und Einflussfaktoren – ES K6 & K7

	Modell 1 B (SE) Beta
Konstante	7.92 (42.18)
Teilkompetenzen	
Phonographisch-silbisches Prinzip	**0.05 (.02) 0.06***
Morphologisches Prinzip	**0.09 (.02) 0.08*** **
Peripheriebereich	**0.37 (.03) 0.36*** **
Prinzip der Wortbildung	**0.24 (.03) 0.26*** **
Wortübergreifendes Prinzip	**0.22 (.02) 0.26*** **
Einflussfaktoren	
Geschlecht[A]	-0.55 (3.27) 0.00
Alter[B]	-1.48 (2.81) -0.01
Migrationshintergrund[C]	-2.86 (4.06) -0.01
Sprache im Haushalt[D]	6.98 (6.72) 0.02
Anzahl der Bücher[E]	-2.40 (1.35) -0.01
Schulform[F]	**16.73 (4.33) 0.08*** **
R^2	**0.945**

N = 255
B (SE) Beta = unstandardisiertes Regressionsgewicht (Standardfehler) standardisiertes Regressionsgewicht
ns = nicht signifikant; * = signifikant (p<.05); ** = signifikant (p<.01) ; *** = signifikant (p<.001)
A = Geschlecht der Schülerinnen und Schüler (0 = weiblich; 1 = männlich)
B = Alter der Schülerinnen und Schüler (offene Angabe)
C = Migrationshintergrund (0 = mindestens ein Elternteil; 1 = kein Elternteil)
D = Sprache im Haushalt (0 = andere Sprache; 1 = Deutsch)
E = Anzahl der Bücher (0 = 100 oder weniger Bücher; 1= 101 und mehr Bücher)
F = Schulform Gymnasium (0 = andere Schulform; 1 = Gymnasium)

Berücksichtigt die Regressionsanalyse sowohl die fünf Teilkompetenzen als auch die sechs Einflussfaktoren zur Erklärung der Kompetenzentwicklung, so zeichnet sich ab, dass die Personenfähigkeit der Schülerinnen und Schüler bei den Teilkompetenzen ausschlaggebend ist und nur durch die Schulform als Einflussfaktor signifikant beeinflusst wird. Demnach scheint die Kompetenzentwicklung von äußeren Einflüssen nur bedingt abhängig zu sein, vielmehr wird sie vorrangig durch den Besuch der jeweiligen Schulform beeinflusst.

Die dargestellten Befunde haben jedoch aufgrund der verhältnismäßig geringen Stichprobengröße nur eine eingeschränkte Aussagekraft. Es ging in dieser Arbeit vor allem darum, Analyseverfahren vergleichend zu untersuchen. Eine Prüfung der Befunde mit den repräsentativen Daten aus den Haupterhebungen in den Klassenstufen 5, 7 und 9 stellt ein zentrales Forschungsdesiderat dar.

8.5 ZUSAMMENFASSENDENDE DISKUSSION UND REFLEXION DER ERGEBNISSE

Aus der Vielzahl der dargestellten Analysen und den daraus resultierenden Ergebnissen wird abschließend eine Gesamtaussage formuliert, die sich als Zusammenfassung und Interpretation versteht.

F1: Wie verhalten sich die in der Entwicklungsstudie und in der Haupterhebung in Klassenstufe 5 ermittelten Gütekriterien für den Rechtschreibtest zueinander?

Die in der NEPS-Studie durchgeführten Entwicklungsstudien (ES) und Haupterhebungen (HE) unterscheiden sich in ihrer Repräsentativität und Stichprobengröße, weshalb im Hinblick auf die Testentwicklung die Aussagekraft der kleineren, nicht repräsentativen Entwicklungsstudien gegenüber den größeren repräsentativen Haupterhebungen zu klären ist. Dies wird mit den Daten der ES und HE der Klassenstufe 5 untersucht (vgl. Kapitel 8.1). Anhand des Studienvergleichs konnte gezeigt werden, dass sich die zur Bewertung der Testentwicklung herangezogenen Item-, Personen- und Modellparameter nur in wenigen bedeutsamen bzw. statistisch nachweisbaren Punkten unterscheiden. Betrachtet man bei beiden Studien die Itemschwierigkeiten der analysierten Wörter und Struktureinheiten des Testinstruments, ist übereinstimmend zu erkennen, dass bei den ganzen Wörtern eine gute Passung zur geschätzten Personenfähigkeit vorliegt und die Struktureinheiten den Schülerinnen und Schülern deutlich leichter fallen als die richtige Schreibweise der ganzen Wörter. Allerdings ist die Annahme des Rasch-Modells gefährdet, da bei Schülerinnen und Schülern mit einer niedrigeren und höheren Personenfähigkeit in beiden Studien unterschiedliche Itemschwierigkeiten festgestellt wurden. Weitergehend ändert sich sowohl bei den ein- und mehrdimensionalen Skalierungen das Ausmaß der statistischen Auffälligkeiten, indem sich bei größeren Stichproben die Item-Fits erhöhen und die Trennschärfe der Items abnimmt. In Bezug auf die geschätzten Personenparameter hat sich gezeigt, dass die Ergebnisse aus der HE K5 im Gegensatz zur ES K5 signifikant höhere Personenfähigkeiten aufweisen, die zur Verletzung der Normalverteilung führen. In Bezug auf die Modellparameter kann anhand der Reliabilitäten eine zuverlässige Erfassung der Rechtschreibkompetenz nachgewiesen werden, die unabhängig von der Stichprobengröße ist. Bei beiden Studien liegen hohe latente Korrelationen innerhalb des differenziellen Kompetenzmodells vor, die zwar je nach Stichprobengröße in ihrer Stärke variieren,

wodurch sich aber die Korrelationsstruktur im Hinblick auf die Eigenständigkeit der Teilkompetenzen nicht ändert. Die Ergebnisse zur Prüfung der Dimensionalität des Kompetenzkonstrukts sind nicht eindeutig und abhängig von der Stichprobengröße, da bei der kleineren Entwicklungsstudie Tendenzen zur Veränderungen der Kompetenzstruktur vorliegen, die weiter erforscht werden müssen.

Dies bedeutet zusammenfassend für die Testentwicklung im Rahmen von Entwicklungsstudien, dass die kleinen Studien geeignet sind, Tests zur kompetenzorientierten Leistungsmessung zu pilotieren, da sie vergleichbare Item-, Personen- und Modellparameter liefern. Unterschiede ergeben sich lediglich im Hinblick auf die Anzahl der statistischen Auffälligkeiten, die bei der Modellierung der Kompetenz auftreten, was zu einer vertretbaren Verringerung der Analysegrundlage führt. Ebenso zeigen sich je nach Studie Unterschiede bei der Dimensionalität des Kompetenzkonstrukts, die jedoch als Vorteil für die Entwicklungsstudien ausgelegt werden können, da sie eine bessere Möglichkeit bieten, die empirische Modellierung des differenziellen Rechtschreibkompetenzmodells explorativ zu erforschen. Zusammenfassend lässt sich feststellen, dass die Testentwicklung in der ES K5 zu einem verlässlichen Instrument führte.

F2: Mit welchen methodischen Verfahren lässt sich die Entwicklung der Rechtschreibkompetenz verlässlich und effizient erfassen?

Zur kompetenzorientierten Leistungsmessung im Längsschnitt sind geeignete Testinstrumente und eine solide methodische Modellierung der Rechtschreibkompetenz erforderlich, um im Zeitverlauf die Entwicklung der Personenfähigkeiten für die einzelnen Schülerinnen und Schüler ermitteln zu können (vgl. Kapitel 8.2). Der aus diesem Grund durchgeführte Vergleich dreier Verfahren zur Modellierung der längsschnittlichen Kompetenzentwicklung hat nach Abwägung der Vor- und Nachteile ergeben, dass je nach Komplexität des zugrundliegenden Kompetenzmodells die getrennte Skalierung oder die Skalierung mit latenten Dimensionen geeignet sein können. Während die Skalierung mit latenten Dimensionen aufgrund der begrenzten Anzahl an Dimensionen nur eingeschränkt einsetzbar ist, stellt die getrennte Skalierung mit latenten Dimensionen die verlässliche und effiziente Voraussetzung dar, um eine längsschnittliche Modellierung und Analyse der Kompetenzentwicklung vorzunehmen. Auf Grundlage der Ergebnisse können für die längsschnittliche Modellierung der Rechtschreibkompetenz je nach

Forschungsinteresse nur die Ankeritems oder auch zusätzliche zeitpunktspezifische Items genutzt werden, da zwischen den Skalierungsvarianten keine signifikanten Unterschiede bestehen. Die Analyse der Kompetenzentwicklung liefert einen besseren Einblick, wenn die Anzahl der Ankeritems in Form von Fixierungsvarianten angepasst wird. Insbesondere aus didaktischer Sicht ermöglichen die Skalierungs- und Fixierungsvarianten die Möglichkeit, den Umgang mit jahrgangsbezogenen Lerninhalten zu betrachten und zu bewerten.

F3: Verändert sich die Kompetenzstruktur von Klassenstufe 6 bis 7, und wenn ja, in welcher Weise?

Die längsschnittliche Modellierung des differenziellen Rechtschreibkompetenzmodells zur Analyse der Kompetenzstruktur liefert auf der Grundlage der Teilkompetenzen tiefergehende Einblicke in die Entwicklung der Rechtschreibkompetenz. Hierbei konnten kleine bis mittlere Entwicklungen der Teilkompetenzen nachgewiesen werden, wobei bei dem wortübergreifenden Prinzip ein mittlerer Effekt und bei den übrigen Teilkompetenzen kleine Effekte vorliegen (vgl. Kapitel 8.3). Folglich ergeben sich differenzielle Kompetenzentwicklungen, die sich anhand der kontrollierten Einflussfaktoren weitergehend erforschen lassen. Demnach entwickeln sich Mädchen besser als Jungen und bei Schülerinnen und Schülern, die altersmäßig über dem für die Klassenstufe üblichen Alter liegen, lässt die Entwicklung in den Teilkompetenzen nach. Mit der Anzahl der im Haushalt zur Verfügung stehenden Bücher verstärkt sich die Kompetenzentwicklung bis zu einem Buchbesitz von 500 Büchern und verringert sich ab 501 Büchern wieder. Weiterhin bestehen Unterschiede je nach besuchter Schulform der Schülerinnen und Schüler, wobei die Gesamtschulen und die Gymnasien die stärksten Unterschiede der differenziellen Kompetenzentwicklungen aufweisen. Keine eindeutigen Unterschiede lassen sich bezüglich eines Migrationshintergrundes und der im Haushalt gesprochenen Sprache nachweisen. Die längsschnittliche Analyse der Kompetenzstruktur liefert somit Hinweise auf eine unterschiedliche Entwicklung der Kompetenzstruktur, die von verschiedenen Faktoren abhängig ist.

Generell ist zu sagen, dass sich die Schülerinnen und Schüler in Abhängigkeit von ihrer sozialen Herkunft und der besuchten Schulform entwickeln. Weitergehend entscheidet die Entwicklung in den Teilkompetenzen über die Kompetenz, ganze Wörter korrekt zu verschriftlichen. Daher bietet die längsschnittliche Analyse der Kompetenzstruktur aus didaktischer Sicht ein großes Forschungspotenzial, um beispielsweise empirisch gesicherte Informationen für die

Auswahl von Lerninhalten in Abhängigkeit vom Lernstand zu gewinnen und gezielte Fördermaßnahmen für die jeweiligen Schülerinnen und Schüler abzuleiten.

F4: Beeinflussen die Entwicklung in den Teilkompetenzen bzw. die Faktoren Geschlecht, Alter, sozioökonomischer Hintergrund, Migrationshintergrund und Schulform die Entwicklung der Rechtschreibkompetenz auf der Ebene des ganzen Wortes von Klassenstufe 6 und 7, und wenn ja, in welcher Weise?

Mögliche Einflüsse auf die Entwicklung der Rechtschreibkompetenz auf der Grundlage des ganzen Wortes wurden mit Korrelations- und Regressionsanalysen untersucht. Zur Erklärung der Kompetenzentwicklung auf der Ebene des ganzen Wortes wurden die Teilkompetenzen und kontrollierte Einflussfaktoren berücksichtigt (vgl. Kapitel 8.4). Auf korrelativer Ebene bestehen hohe Zusammenhänge zwischen dem ganzen Wort und den Teilkompetenzen, die sich auch in der Regressionsanalyse in Form von signifikanten Effekten widerspiegeln. Demnach tragen die Teilkompetenzen mit einer zu knapp 95 Prozent aufgeklärten Varianz zur Kompetenzentwicklung beim ganzen Wort bei. Die kontrollierten Einflussfaktoren bestätigen anhand der korrelativen Befunde für das Geschlecht, das Alter, die Anzahl der Bücher und die Schulform einen Zusammenhang zur Kompetenzentwicklung auf der Ebene des ganzen Wortes, wobei sich in der multiplen Regressionsanalyse nur das Geschlecht und die Schulform signifikant durchsetzen konnten. Die Regressionsanalyse mit den Teilkompetenzen und den Einflussfaktoren liefert das Ergebnis, dass sich in der multivariaten Betrachtung nur die Teilkompetenzen und die Schulform als prädiktiv für die Entwicklung der Rechtschreibkompetenz beim ganzen Wort erweisen.

Daraus lässt sich folgern, dass die Entwicklung der Rechtschreibkompetenz in der Sekundarstufe I weniger von der sozialen Herkunft abhängig ist, sondern vielmehr in Zusammenhang mit der Beherrschung der Teilkompetenzen steht, wobei die Kompetenz im Peripheriebereich, bei der Wortbildung und beim wortübergreifenden Prinzip den Ausschlag geben. Allein die Schulform hat darüber hinaus einen nachweislichen Einfluss auf den Verlauf der Kompetenzentwicklung.

9. FAZIT UND AUSBLICK

In diesem Kapitel werden die theoretischen, methodischen und empirischen Zugänge dieser Arbeit aufgegriffen, multikriterial in den Stand der Kompetenzforschung eingeordnet und zusammengeführt. Zum Abschluss erfolgt ein Ausblick, in dem weitergehende Zielsetzungen zur quantitativen Datenanalyse der längsschnittlichen Erfassung der Rechtschreibkompetenz skizziert werden.

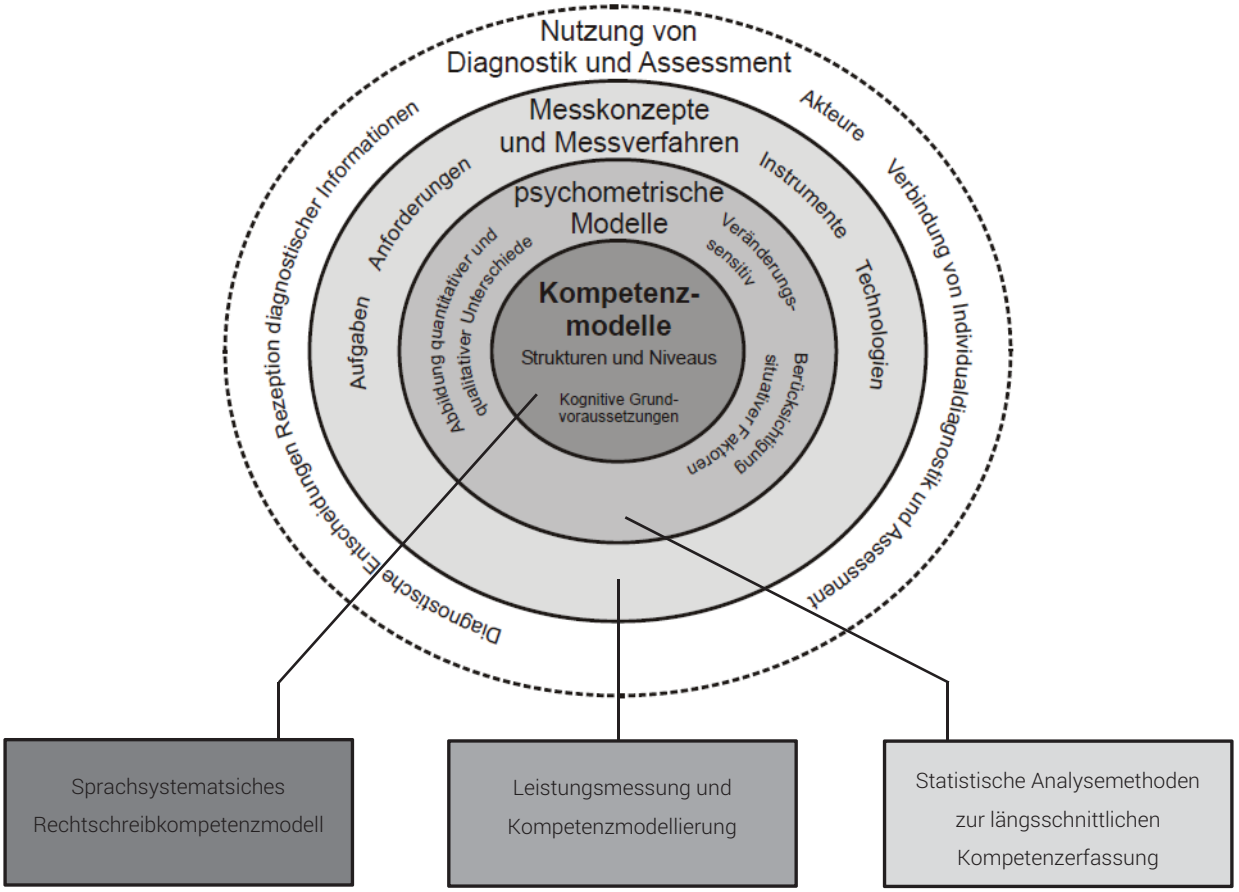
Die Einordnung der verschiedenen Zugänge dieser Arbeit erfolgt anhand des Schwerpunktprogramms „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (Kapitel 2.1.2.2), indem die Arbeitsbereiche des Schwerpunktprogramms (SPP) aufgegriffen und zur Darstellung der Zielaspekte der quantitativen Datenanalyse zur längsschnittlichen Erfassung der Rechtschreibkompetenz genutzt werden.

Verfolgt man den aktuellen Trend der empirischen Bildungsforschung (Kapitel 1), so sind internationale und nationale Leistungsvergleiche im Rahmen von Bildungspanel bzw. Längsschnittstudien State of the Art. Zudem zeigt sich eine Abwendung von der reinen Leistungsmessung hin zu einer kompetenzorientierten Leistungsmessung mit dem Ziel, eine Kompetenzmodellierung vorzunehmen (Kapitel 2). Diese basiert, wie es der Forschungsstand fordert, auf einem theoretisch begründeten Kompetenzmodell, das mithilfe von psychometrischen Verfahren modelliert und empirisch validiert wird. Dafür sind geeignete Messkonzepte und -verfahren notwendig (Kapitel 3). Nur wenige Testinstrumente zur kompetenzorientierten Leistungsmessung der Rechtschreibung eignen sich jedoch unter Berücksichtigung der Standards für die längsschnittlichen Vergleichsstudien (Kapitel 1.2.3) und die kompetenzorientierte Leistungsmessung (Kapitel 2.3.3) für Diagnostik und Assessment. Der sprachsystematische Rechtschreibtest (SRT) (Kapitel 5.2) erfüllt diese Anforderung als Testinstrument zur Erfassung der Rechtschreibkompetenz, da das zugrundeliegende Kompetenzmodell den fachdidaktischen Theorie- und Forschungsstand berücksichtigt und eine evidenzbasierte Überprüfung vorliegt.

Die aufgezeigten theoretischen, methodischen und empirischen Zugänge dieser Arbeit in Form der Leistungsmessung und Kompetenzmodellierung (Kapitel 2), die dazugehörigen statistischen Analysemethoden zur längsschnittlichen Kompetenzerfassung (Kapitel 3) und das zugrundeliegende sprachsystematische Rechtschreibkompetenzmodell (Kapitel 5) werden mit

der Zuordnung zu den Arbeitsbereichen des Schwerpunktprogramms in Abbildung 9.1 überblicksartig dargestellt und nachfolgend näher ausgeführt.

Abbildung 9.1: Arbeitsbereiche des Schwerpunktprogramms am Beispiel der quantitativen Datenanalyse zur längsschnittlichen Erfassung der Rechtschreibkompetenz



Im Mittelpunkt des Schwerpunktprogramms stehen die Kompetenzmodelle. Dem wird in dieser Arbeit durch das sprachsystematische Rechtschreibkompetenzmodell (vgl. Kapitel 5) Rechnung getragen. Es bildet im Einklang mit der Forschung zur Rechtschreibung die theoretische Grundlage ab (vgl. Blatt et al., 2011; Voss et al., 2007). So gibt es Belege, dass der herkömmliche Rechtschreibunterricht zu großen Verunsicherungen bei den Schülerinnen und Schülern führt (vgl. Kapitel 0). Dies wird von Vertretern der Graphematik darauf zurückgeführt, dass die dem herkömmlichen Unterricht zugrundeliegende Vorstellung der Schriftsprache als Abbild der mündlichen Sprache dazu führt, dass die Schriftsprache als undurchschaubares System einer Vielzahl von Regeln und Ausnahmen vermittelt wird (z. B. Eisenberg & Fuhrhop,

2007). Dass die Orthografie nicht nur beherrscht, sondern auch verstanden werden kann, und das auf der Verstehensgrundlage ein selbständiges Lernen ermöglicht wird, legen die Ergebnisse der graphematischen Forschung nahe, auf deren Basis das sprachsystematische Kompetenzmodell entwickelt wurde (Kapitel 5). Diese Ergebnisse stellen aus aktueller didaktischer Sicht eine geeignete Grundlage für die Auswahl der Lerninhalte bereit. Darüber hinaus kann mittels des sprachsystematischen Rechtschreibtests (vgl. Kapitel 5.2) das Lernergebnis der Schülerinnen und Schüler differenziell untersucht werden. Dabei wird berücksichtigt, dass die Rechtschreibkompetenz keine globale Kompetenz darstellt, sondern aus Teilkompetenzen besteht. Die Ausprägung der Kompetenz in den Teilkompetenzen, also die Höhe und das Verhältnis zueinander, entscheidet über die Fähigkeit der Schülerinnen und Schüler, Wörter orthografisch korrekt zu schreiben. Dabei sind die neuen Lerninhalte in der Sekundarstufe I entscheidend, weshalb dem Rechtschreibunterricht in der Sekundarstufe I eine große Bedeutung zugesprochen wird. Ebenso ist der sprachsystematische Rechtschreibtest wegen der differenziellen Erfassung der Rechtschreibkompetenz als Assessment-Instrument als Grundlage für ein individualisiertes Lernen geeignet. Dies setzt allerdings computerisierte Verfahren voraus (vgl. Frahm 2013).

Die methodischen und empirischen Grundlagen dieser Arbeit stellen eine Umsetzung der beiden Arbeitsbereiche „psychometrische Modelle“ und „Messkonzepte und Messverfahren“ des Schwerpunktprogramms dar, indem die Leistungsmessung und Kompetenzmodellierung (vgl. Kapitel 2) inklusive der dazu notwendigen statistischen Analysemethoden zur längsschnittlichen Kompetenzerfassung (vgl. Kapitel 3) aufgearbeitet werden. Dies umfasst die Auseinandersetzung mit den Begriffen Leistung und Kompetenz, um daraufhin die kompetenzorientierte Leistungsmessung abzuleiten bzw. zu definieren. Dazu gehören die Kompetenzmessung und -modellierung anhand eines psychometrischen Modells. Die Umsetzung zur Messung und Modellierung des psychometrischen Modells erfordern geeignete Messkonzepte und -verfahren, die im Rahmen dieser Arbeit in Form von statistischen Analysemethoden zur längsschnittlichen Kompetenzerfassung umgesetzt werden. Zentral ist hierfür die Verwendung der Item-Response-Theory zur latenten Modellierung der Rechtschreibkompetenz. Um dem Ziel der längsschnittlichen Kompetenzerfassung gerecht zu werden, ist die Berücksichtigung von Modellen zur Veränderungsmessung notwendig, um die längsschnittliche Entwicklung der Rechtschreibkompetenz zu messen und zu modellieren.

Über diese Arbeit hinaus sind in Bezug auf das sprachsystematische Rechtschreibkompetenzmodell im Rahmen der NEPS-Studie entsprechende Arbeiten zur Unterrichtsforschung zu finden, die eine praxisbezogene Anwendung in Form der „Nutzung von Diagnostik und Assessment" des Schwerpunktprogramms realisieren (vgl. Frahm, 2013; Prosch, in Vorb.). Es handelt sich dabei um dafür notwendige Grundlagenforschung.

Die veranschaulichte Einordnung der theoretischen, methodischen und empirischen Zugänge dieser Arbeit in die Arbeitsbereiche des Schwerpunktprogramms erweist sich als eine sinnvolle Zusammenführung, da die geforderten Zielaspekte des SPPs ebenfalls im Kontext der quantitativen Datenanalyse zur längsschnittlichen Erfassung der Rechtschreibkompetenz vorzufinden sind. Daher kann diese Arbeit als ein Forschungsbeispiel bezeichnet werden, das sich an aktuellen Forschungsdesideraten orientiert und einen Beitrag dazu liefert, theoretische Kompetenzmodelle für kognitive Kompetenzen mit psychometrischen Modellen und Messverfahren adäquat zu verknüpfen und die Kompetenzmodelle als Grundlage für pädagogische Entscheidungen zu nutzen.

Abschließend werden in einem Ausblick die Standards im Bereich der längsschnittlichen Vergleichsstudien (Kapitel 1.2.3) und der Kompetenzmessung (Kapitel 2.3.3) aufgegriffen, um weiteren Forschungsbedarf auf der Grundlage der empirischen Ergebnisse (vgl. Kapitel 8) zu benennen.

Hinsichtlich der Testentwicklung wurde aufgezeigt, dass die verhältnismäßig kleinen Entwicklungsstudien zur Erprobung des sprachsystematischen Rechtschreibtests geeignet sind und sich die Aussagekraft im Vergleich zu den Untersuchungen mit den Haupterhebungsdaten nur unwesentlich unterscheidet. Da diese Überprüfung lediglich für die Klassenstufe 5 stattfand, ist es ratsam, die Daten der Entwicklungsstudie und Haupterhebung der Klassenstufe 7 als Kontrolle heranzuziehen und die Ergebnisse bezüglich der Aussagekraft zur Testentwicklung zu replizieren. Dies hat nicht nur den Zweck, die Ergebnisse abzusichern, sondern kann auch zur Optimierung des Testinstruments bzw. der Auswertungsverfahren sowie der Rückmeldung von Ergebnissen genutzt werden.

Die in dieser Arbeit verglichenen Auswertungsverfahren zur Längsschnittforschung sind essenziell für die adäquate Modellierung der Rechtschreibkompetenz im Rahmen der NEPS Studie, weshalb sich ein erneuter Vergleich der drei Verfahren zur längsschnittlichen Kompe-

tenzmodellierung für personenspezifische Veränderungen mit den Stichproben der Haupterhebungen in den Klassenstufen 5 und 7 anbietet.

Die Forschung zur Kompetenzstruktur der Rechtschreibung liefert aktuell nur Hinweise für die unterschiedliche differenzielle Kompetenzentwicklung der Schülerinnen und Schüler, weshalb die Ausweitung der Untersuchung sowohl aus statistischer als auch didaktischer Sicht erstrebenswert ist. Neben der deskriptiven und induktiven Betrachtungsweise der Kompetenzstruktur ist zudem die Ermittlung von Kompetenzprofilen wichtig, um die Entwicklung der Rechtschreibkompetenz weitergehend zu abstrahieren und der didaktischen Forschung zugänglich zu machen. Auf der Grundlage von Kompetenzprofilen können Rechtschreibprobleme von Schülerinnen und Schülern mit unterschiedlichem Kompetenzstand identifiziert und beschrieben werden, um gezielte Fördermaßnahmen oder auch effiziente Unterrichtskonzepte zu entwickeln. Gerade der tiefergehenden Auseinandersetzung mit der längsschnittlichen Veränderung der Kompetenzstruktur ist bislang nicht genügend nachgegangen worden. Sie bildet aber eine Voraussetzung dafür, abgesicherte Aussagen zur differenziellen Kompetenzentwicklung treffen zu können.

Weitergehend ist der Einbezug von Einflussfaktoren auf die Kompetenz in den Forschungsprozess zur kompetenzorientierten Leistungsmessung wichtig. Vielversprechend erscheint z. B. die Kontrolle des „Rechtschreib-Engagements“, das im Schülerfragebogen erfasst wurde. Mit entsprechenden Analysen können die Bedingungen der Kompetenzentwicklung der Schülerinnen und Schüler erkundet werden, die sowohl der Testentwicklung als auch der Unterrichtsforschung zugutekommen. Dazu eignen sich multivariate Analysemethoden, die das Zusammenspiel mehrerer Faktoren in Bezug auf die längsschnittliche Kompetenzentwicklung der Rechtschreibung aufzeigen können. Darüber hinaus gilt es, den quantitativen Zugang zur längsschnittlichen Entwicklung der Rechtschreibkompetenz um qualitative Analysen zu erweitern, weshalb die Fachdidaktik bei der Begutachtung und Interpretation der quantitativen Ergebnisse substanziell einbezogen werden muss. Die Triangulation der quantitativen und qualitativen Forschung zur Rechtschreibkompetenz wird in der Scientific Community zur Diskussion gestellt (vgl. Blatt, Frahm, Jarsinski & Prosch, 2013). Mithilfe der skizzierten Erkenntnisse kann die quantitative Datenanalyse zur längsschnittlichen Erfassung der Rechtschreibkompetenz weiter optimiert und in zukünftigen Studien erneut auf ihre Evidenz hin überprüft werden.

10. LITERATURVERZEICHNIS

Adams, R. J. & Carstensen, C. H. (2002). Scaling outcomes. In R. J. Adams & M. Wu (Hrsg.), *PISA 2000 Technical Report* (S. 149–162). Paris: OECD.

Adams, R. J. (2002). Scaling PISA cognitive data. In R. Adams & M. Wu (Hrsg.), *PISA 2000 technical report* (S. 99–108). Paris: OECD.

Allmendinger, J., Ebner, C. & Nikolai, R. (2010). Soziologische Bildungsforschung. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (S. 47–70). Wiesbaden: VS Verlag für Sozialwissenschaften.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50* (1), 3–16.

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S. & Blossfeld, H.-P. (2011). Sampling Design of the National Educational Panel Study: Challenges and Solutions. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Eds.), *Education as a Lifelong Process. The German National Educational Panel Study (NEPS)* (Zeitschrift für Erziehungswissenschaft, Special Issues 14, S. 51–65). Wiesbaden: VS Verlag für Sozialwissenschaften.

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2008). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung* (12. Aufl.). Berlin, Heidelberg: Springer Verlag.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J. & Weiß, M. (2001). *PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Blatt, I. (2010). Sprachsystematische Rechtschreibdidaktik: Konzept, Materialien, Tests. In H. Bredel, A. Müller & G. Hinney (Hrsg.), *Schriftkompetenz und Schriftsystem: linguistisch, empirisch, didaktisch* (S. 101–132). Tübingen: Niemeyer Verlag.

Blatt, I. & Frahm, S. (2013). Explorative Analysen zur Entwicklung der Rechtschreibkompetenz im Rahmen der NEPS-Studie (Klassenstufe 5–7). *Didaktik Deutsch, 34*, 12–36.

Blatt, I., Frahm, S., Jarsinski, S. & Prosch, A. (2013, September). *Self-Concept, Motivation, Strategies and Instruction – Exploring Determinants of Spelling Competence from a Multidimensional Perspective*. ECER 2013, Istanbul.

Blatt, I., Frahm, S., Prosch, A., Jarsinski, S. & Voss A. (in Vorb.). *Kompetenzmodellierung im Kontext des Nationalen Bildungspanels (NEPS) am Beispiel der Rechtschreibkompetenz*. Münster: Waxmann.

Blatt, I. & Jarsinski, S. (2009). Sprachsystematische Rechtschreibdidaktik als Fördergrundlage: Ein Fallbeispiel. In R. Valtin & B. Hofmann (Hrsg.), *Kompetenzmodelle der Orthographie. Empirische Befunde und förderdiagnostische Möglichkeiten* (S. 91–112). Berlin: DGLS.

Blatt, I., Müller, A. & Voss, A. (2010). Schriftstruktur als Lesehilfe. Konzeption und Ergebnisse eines Hamburger Leseförderprojekts in Klasse 5 (HeLp). In G. Hinney (Hrsg.), *Schriftsystem und Schrifterwerb: linguistisch – didaktisch – empirisch* (Reihe Germanistische Linguistik, Bd. 289, 1. Aufl., S. 171–202). Berlin: Walter de Gruyter GmbH Co. KG.

Blatt, I., Ramm, G. & Voss, A. (2009). Modellierung und Messung der Textkompetenz im Rahmen einer Lernstandserhebung in Klasse 6. *Didaktik Deutsch, 26*, 54–81.

Blatt, I., Voss, A., Kowalski, K. & Jarsinski, S. (2011). Messung von Rechtschreibleistung und empirische Kompetenzmodellierung. In U. Bredel & T. Reißig (Hrsg.), *Weiterführender Orthographieerwerb* (S. 226–256). Baltmannsweiler: Schneider Verlag Hohengehren.

Blossfeld, H.-P., Schneider, T. & Doll, J. (2009). Methodological Advantages of Panel Studies: Designing the New National Educational Panel Study (NEPS) in Germany. *Journal for Educational Research Online, 1* (1), 10–32. Retrieved May 15, 2014, from http://www.j-e-r-o.com/index.php/jero/article/view/59/47

Blossfeld, H.-P., von Maurice, J. & Schneider, T. (2011a). *Grundidee, Konzeption und Design des Nationalen Bildungspanels für Deutschland (NEPS Working Paper No. 1).* Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Blossfeld, H.-P., von Maurice, J. & Schneider, T. (2011b). The National Educational Panel Study: need, main features, and research potential. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Eds.), *Education as a Lifelong Process. The German National Educational Panel Study (NEPS)* (Zeitschrift für Erziehungswissenschaft, Special Issues 14, S. 5–17). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6* (4), 431–444.

Böhme, K. & Bremerich-Vos, A. (2009). Diagnostik der Rechtschreibkompetenz in der Grundschule – Konstruktprüfung mittels Fehler- und Dimensionsanalysen. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 330-356). Weinheim: Beltz.

Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2004) The concept of validity. *Psychological Review, 111* (4), 1061–1071.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Sozialwissenschaftler* (4. Aufl.). Heidelberg: Springer.

Bortz, J. (2004). *Statistik für Human- und Sozialwissenschaftler*. Heidelberg: Springer.

Bos, W., Bonsen, M. & Gröhlich, C. (2009). *KESS 7 – Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7*. Münster: Waxmann.

Bos, W., Bremerich-Vos, A., Tarelli, I. & Valtin, R. (2012). Lesekompetenzen im internationalen Vergleich. In W. Bos, I. Tarelli, A. Bremerich-Vos & K. Schwippert (Hrsg.), *IGLU 2011 – Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 91–135). Münster: Waxmann.

Bos, W. & Gröhlich, C. (2009). Special Issue Editorial: Longitudinal Assesments and Panel Studies in Educational Research. *Journal for Educational Research Online, 1* (1), 7–9. Retrieved May 15, 2014, from http://www.j-e-r-o.com/index.php/jero/article/view/58/46

Bos, W. & Gröhlich, C. (2010). *KESS 8 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 8*. Münster: Waxmann.

Bos, W., Holtappels, H. G. & Rösner, E. (2006). Schulinspektion in den deutschen Bundesländern – eine Baustellenbeschreibung. In W. Bos, H. G. Holtappels, H. Pfeiffer, H.-G. Rolff

& R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung Band 14 – Daten, Beispiele und Perspektiven* (S. 81–125). Weinheim: Juventa.

Bos, W., Lankes, E. M., Prenzel, M., Schwippert, K., Walther, G. & Valtin, R. (2003). *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich*. Münster: Waxmann.

Bos, W. & Pietsch, M. (2006). *KESS 4 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen*. Münster: Waxmann.

Bos, W., Schwippert, K. & Stubbe, T. C. (2007). Die Koppelung von sozialer Herkunft und Schülerleistung im internationalen Vergleich. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, K. Schwippert & R. Valtin (Hrsg.), *IGLU 2006 – Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 225–247). Münster: Waxmann.

Bos, W., Tarelli, I., Bremerich-Vos, A. & Schwippert, K. (2012). IGLU 2011 – Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich. Münster: Waxmann.

Bos, W., Wendt, H., Köller, O. & Selter, C. (2012). *TIMSS 2011 – Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

Bourdieu, Pierre (1983). Ökonomisches Kapital – Kulturelles Kapital – Soziales Kapital. In R. Kreckel (Hrsg.), *Soziale Ungleichheiten* (Sozialen Welt, Sonderband 2, S. 183–198). Göttingen: Schwartz.

Brehl, T., Wendt, H. & Bos, W. (2012). Geschlechtsspezifische Unterschiede in mathematischen und naturwissenschaftlichen Kompetenzen. In W. Bos, H. Wendt, O. Köller & C. Selter (Hrsg.), *TIMSS 2011 – Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 203–230). Münster: Waxmann.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson.

Bulheller, S., Ibrahimovic, N. & Häcker, H. O. *(2005). RST-NRR. Rechtschreibtest – Neue Rechtschreiberegelung* (2. Aufl.). Frankfurt am Main: Harcourt Test Services.

Carstensen, C. H., Lankes, E.-M. & Steffensky, M. (2012). Modellierung von längsschnittlichen Daten am Beispiel einer quasi-experimentellen Studie zur Erfassung von naturwissenschaftlichen Kompetenzen im Kindergartenalter. In W. Kempf & R. Langeheine (Hrsg.), *Item-Response-Modelle in der sozialwissenschaftlichen Forschung* (S. 109–126). Berlin: Regener Verlag.

Chomsky, C. (1969). *The Acquisition of Syntax in Children from 5 to 10*. Cambridge: Mass.

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

Coleman, J. S. (1966). *Equality of educational opportunity*. Washington, D.C.: U.S. Department of Health, Education, and Welfare.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology, 94 (Supplement*), 95–120.

Dahrendorf, R. (1965). *Bildung ist Bürgerrecht. Plädoyer für eine aktive Bildungspolitik*. Hamburg: Nannen.

Deutscher Bildungsrat (1974). *Zur Neuordnung der Sekundarstufe II. Konzept für eine Verbindung von allgemeinem und beruflichem Lernen*. Bonn: Klett.

Diekmann, A. (2007). *Empirische Sozialforschung. Grundlage, Methoden, Anwendungen*. Reinbek: Rowohlt Taschenbuch Verlag GmbH.

Ditton, H. (2011). Entwicklungslinien der Bildungsforschung. Vom deutschen Bildungsrat zu aktuellen Themen. In H. Reinders, H. Ditton, C. Gräsel, & B. Gniewosz, (Hrsg.), *Empirische Bildungsforschung. Strukturen und Methoden* (Bd. 1, S. 29–42). Wiesbaden: VS Verlag für Sozialwissenschaften.

Ditton, H. & Reinders, H. (2011). Überblick: Felder der Bildungsforschung. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Lehrbuch Empirische Bildungsforschung – Gegenstandsbereiche* (S. 69–74). Wiesbaden: VS Verlag für Sozialwissenschaften.

Drechsel, B., Prenzel, M. & Seidel, T. (2009). Nationale und internationale Schulleistungsstudien. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 354–380). Heidelberg: Springer.

Duden, K. (1902). *Orthographisches Wörterbuch der deutschen Sprache* (7. Aufl.). Leipzig: Bibliographisches Institut.

Duller, C. (2006). *Einführung in die Statistik mit EXCEL und SPSS. Ein anwendungsorientiertes Lehr- und Arbeitsbuch*. Heidelberg: Physica-Verlag.

Dunkin, M. J. & Biddle, B. J. (1974). *The study of teaching*. New York: Holt and Rinehart.

Edding, F. (1963). *Ökonomie des Bildungswesens. Lehren und Lernen als Haushalt und als Investition* (1. Aufl.). Freiburg: Rombach.

Eggert, S., Bögeholz, S., Watermann, R. & Hasselhorn, M. (2010). Förderung von Bewertungskompetenz im Biologieunterricht durch zusätzliche Strukturierungshilfen beim Kooperativen Lernen – Ein Beispiel für Veränderungsmessung. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 299–314.

Eisenberg, P. (1995). Der Buchstabe und die Schriftstruktur des Wortes. In Duden (Hrsg.), *Die Grammatik. Band 4* (S. 56–84). Mannheim: Dudenverlag.

Eisenberg, P. (2004). *Grundriss der deutschen Grammatik, Bd. 1: Das Wort* (2. überarb. und aktual. Aufl.). Stuttgart: Metzler.

Eisenberg, P. & Fuhrhop, N. (2007). Schulorthographie und Graphematik. *Zeitschrift für Sprachwissenschaft, 26*, 15–41.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and environment. *Psychometrika, 56* (3), 495–515.

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

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

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

Flechsig, K.-H., Tütken, H., Riedel, K., Thiersch, H. & Skowronek, H. (1968). Die Steuerung und Steigerung der Lernleistung durch die Schule. In H. Roth (Hrsg.), *Begabung und Lernen* (Bd. 4, S. 449–503). Stuttgart: Klett.

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

Frahm, S. & Blatt, I. (2011). Rechtschreibtests. In U. Bredel (Hrsg.), *Weiterführender Orthographieunterricht* (S. 546–570). Baltmannweiler: Schneider Verlag Hohengehren.

Frahm, S., Goy, M., Kowalski, K., Sixt, M., Strietholt, R., Blatt, I., Bos, W. & Kanders, M. (2011). 14 Transition and development from lower secondary to upper secondary school. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Eds.), *Education as a Lifelong Process. The German National Educational Panel Study (NEPS)* (Zeitschrift für Erziehungswissenschaft, Special Issues 14, S. 217–232). Wiesbaden: VS Verlag für Sozialwissenschaften.

Frahm, S. & Jarsinski, S. (2010, August). *Modelling Longitudinal Orthography Data with IRT – Results of an Intervention Control Study in Germany (HeLp 2007/08)*. ECER 2010, Helsinki.

Frahm, S. & Jarsinski, S. (2011, September). *Language Systematic Tests (paper/pencil and computerized) for Evaluating the Development of Spelling Competency*. ECER 2011, Berlin.

Frey, A. (2012). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 275–293). Berlin: Springer Medizin Verlag.

Fricke, R. (1974). *Kriteriumsorientierte Leistungsmessung*. Stuttgart: Kohlhammer.

Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. Patterson, J. Marshall, & M. Coltheart (Eds.), *Surface Dyslexia, Neuropsychological and Cognitive Studies of Phonological Reading* (pp 301–330). London: Erlbaum.

Fuhrhop, N. (2006). *Orthografie*. Heidelberg.

Gläser-Zikuda, M. (2011). Qualitative Auswertungsverfahren. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), Empirische Bildungsforschung. Strukturen und Methoden (Bd. 1, S. 109–119). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gniewosz, B. (2011a). Kompetenzentwicklung. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Empirische Bildungsforschung* (Bd. 2, S. 57–67). Wiesbaden: VS Verlag.

Gniewosz, B. (2011b). Testverfahren. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Empirische Bildungsforschung* (Bd. 1, S. 68–76). Wiesbaden: VS Verlag.

Goldhammer, F. & Hartig, J. (2012). Interpretation von Testresultaten und Testeichung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 165–192). Berlin: Springer.

Gräsel, C. (2011). Was ist Empirische Bildungsforschung? In H. Reinders, H. Ditton, C., Gräsel, & B. Gniewosz (Hrsg.), *Empirische Bildungsforschung – Strukturen und Methoden* (S. 13–-28). Wiesbaden: VS Verlag für Sozialwissenschaften.

Grund, M., Haug, G. & Naumann, C. L. (2004). *Diagnostischer Rechtschreibtest für 5. Klassen: DRT 5*. Göttingen: Beltz Verlag.

Häder, M. (2010). *Empirische Sozialforschung*. Wiesbaden: VS Verlag.

Hagenaars, J. A. & McCutcheon, A. L. (2002). *Applied Latent Class Analysis*. Cambridge: Cambridge University Press.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Harring, M., Rohlfs, C. & Palentien, C. (2007). *Perspektiven der Bildung*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Hartig, J. (2008). Kompetenzen als Ergebnisse von Bildungsprozessen. In N. Jude, J. Hartig & E. Klieme (Hrsg.), *Kompetenzerfassung in pädagogischen Handlungsfeldern. Theorien, Konzepte und Methoden* (S. 15–25). Berlin: BMBF.

Hartig, J. & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In J. Hartig & E. Klieme (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (S. 17–36). Bonn: Bundesministerium für Bildung und Forschung (BMBF).

Hartig, J., Jude, N. & Wagner, W. (2008). Methodische Grundlagen der Messung und Erklärung sprachlicher Kompetenzen. In E. Klieme, W. Eichler, A. Helmke, R. H. Lehmann, G. Nold, H.-G. Rolff, K. Schröder, G. Thomé, & H. Willenberg (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch* (S. 34–54). Weinheim: Beltz.

Hartig, J. & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127–143). Berlin: Springer.

Hartig, J. & Kühnbach, O. (2006). Schätzung von Veränderung mit „plausible values" in mehrdimensionalen Rasch-Modellen. In A. Ittel & H. Merkens (Hrsg.), *Veränderungsmessung und Längsschnittstudien in der empirischen Erziehungswissenschaft* (S. 27–44). Wiesbaden: VS Verlag für Sozialwissenschaften.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19* (1), 49–78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9* (2), 139–164.

Hattie, J. (2009). *Visible Learning. A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Herné, K. L. & Naumann, C. L. (2002). *Aachener Förderdiagnostische Rechtschreibfehler-Analyse – AFRA. Systematische Einführung in die Praxis der Fehleranalyse* (4. Aufl.). Aachen: Alpha Zentaurus Verlag.

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

Hopf, C. (2004). *Die experimentelle Pädagogik. Empirische Erziehungswissenschaft in Deutschland am Anfang des 20. Jahrhunderts*. Bad Heilbrunn: Klinkhardt.

Hornberg, S. & Bos, W. (2007). Schule im internationalen Vergleich: Der Beitrag von internationalen Schulleistungsstudien am Beispiel von PIRLS/IGLU. In M. Harring, C. Rohls & C. Palentien (Hrsg.), *Perspektiven der Bildung – Kinder und Jugendliche in formellen, nichtformellen und informellen Bildungsprozessen* (S. 155–183). Wiesbaden: VS Verlag für Sozialwissenschaften.

Hurrelmann, K. (2002). Selbstsozialisation oder Selbstorganisation? *Zeitschrift für Soziologie der Erziehung und Sozialisation, 22* (2), 155–166.

Jarsinski, S. & Frahm, S. (2012). *The Longitudinal data collection of spelling competence in the National education panel study (NEPS): Test evaluation and data modeling*. ECER 2012, Cadiz.

Jarsinski, S., Frahm, S., Blatt, I., Bos, W. & Kanders, M. (subm.). Assessing spelling competence in the longitudinal study National Educational Panel Study – stage-specific supplement. In H.-P. Blossfeld, J. von Maurice & J. Skopek (Eds.), *Methodological Issues of Longitudinal Surveys: The Example of the National Educational Panel Study*.

Jude, N. & Klieme, E. (2008). Einleitung. In N. Jude, J. Hartig & E. Klieme (Hrsg.), *Kompetenzerfassung in pädagogischen Handlungsfeldern. Theorien, Konzepte und Methoden* (Bildung – Ideen zünden!; S. 9–13). Bonn: BMBF.

Kendall, M. G. & Stuart, A. (1973). *The advanced theory of statistics* (Vol. 2). New York: Hafner.

Kersting, M. & Althoff, K. (2004). *Rechtschreibungstest (RT)*. Göttingen: Hogrefe.

Klieme, E. (2009). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise (Bildung – Ideen zünden!)*. Bonn: BMBF.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E. & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise. Stand Juni 2003* (Bildungsreform 1). Bonn: BMBF.

Klieme, E. & Hartig, J. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. In M. Prenzel, I. Gogolin & H.-H. Krüger (Hrsg.), *Kompetenzdiagnostik* (Zeitschrift für Erziehungswissenschaft, Sonderheft. 8, S. 11–29). Wiesbaden: VS Verlag für Sozialwissenschaften.

Klieme, E. & Leutner, D. (2006a). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Überarbeitete Fassung des Antrags an die DFG auf die Einrichtung eines Schwerpunktprogramms. Zugriff am 15.05.2014. Verfügbar unter: http://kompetenzmodelle.dipf.de/pdf/rahmenantrag

Klieme, E. & Leutner, D. (2006b). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingrichteten 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. In J. Hartig & E. Klieme (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (S. 1–15). Bonn: BMBF.

KMK (2012). *Vereinbarung zur Weiterentwicklung von VERA*. Beschluss der KMK vom 08.03.2012. Zugriff am 15.05.2014. Verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2012/2012_03_08_Weiterentwicklung-VERA.pdf

Köhler, P. (2008). ANDREE, Analyse der Rechtschreibentwicklung, Programm zur Feststellung der Rechtschreibfähigkeiten von fortgeschrittenen Schreiberinnen und Schreibern. Braunschweig: Köhler.

Kolen, M. J. & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice, 29* (3), 8–14.

Kolen, M. J., Tong, Y., & Brennan, R. L. (2011). Scoring and scaling educational tests. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 43–58). New York: Springer.

Köller, O. (2012). What works best in school? Hatties Befunde zu Effekten von Schul- und Unterrichtsvariablen auf Schulleistungen. *Psychologie in Erziehung und Unterricht, 59* (1), 72–78.

Kormann, A. & Horn, R. (2006). *SSB – Screeningverfahren für Schul- und Bildungsberatung. Rechtschreiben Jahrgangsstufe 1 bis 10 und Intelligenz Jahrgangsstufen 4 bis 10.* Frankfurt am Main: Swets test Service.

Kosog, O. (1912). *Unsere Rechtschreibung und die Notwendigkeit ihrer gründlichen Reform*. Leipzig: Teubner.

Kristen, C., Römmer, A., Müller, W. & Kalter, F. (2005). *Längsschnittstudien für die Bildungsberichterstattung – Beispiele aus Europa und Nordamerika*. Bonn: BMBF.

Lay, W. A. (1918). *Experimentelle Pädagogik. Mit besonderer Rücksicht auf die Erziehung durch die Tat* (Bd. 224). Leipzig: Teubner.

Löffler, I. & Meyer-Schepers, U. (2001). *Dortmunder-Schriftkompetenz-Ermittlung. DoSE*. Unveröffentlichtes Manuskript, Dortmund.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Marx, H. (1999). Rechtschreibleistung vor und nach der Rechtschreibreform: Was ändert sich bei Grundschulkindern? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 31* (4), 180–189.

May, P. (2001). *Hamburger Schreib-Probe: HSP; zur Erfassung der grundlegenden Rechtschreibstrategien*. Hamburg: Verlag für pädagogische Medien.

Meiser, T. (2007). Rasch models for longitudinal data. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models. Extensions and applications* (pp. 191–199). New York: Springer.

Meumann, E. (1914). Abriss der experimentellen Pädagogik. Leipzig: Engelmann.

Mislevy, R. J., Beaton, A., Kaplan, B. A., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29* (2), 133–161.

Moosbrugger, H. (2012). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 227–274). Berlin: Springer Medizin Verlag.

Moosbrugger, H. & Kelava, A. (2012). Testtheorie und Fragebogenkonstruktion (2. Aufl.). Berlin: Springer Medizin Verlag.

Naumann, C. L. (1999). *Orientierungswortschatz – Die wichtigsten Wörter und Regeln für die Rechtschreibung Klassen 1–6*. Weinheim: Beltz.

OECD (2005). *PISA 2003 Technical Report*. Paris: OECD.

Paek, I. (2002). *Investigation of differential item function: Comparisons among approaches, and extension to a multidimensional context*. Unpublished PhD Dissertation. Berkeley, CA: Universtity of California.

Picht, G. (1964). *Die deutsche Bildungskatastrophe. Analyse und Dokumentation*. Olten: Walter-Verlag.

Prenzel, M., Sälzer, C., Klieme, E. & Köller, O. (2013). *PISA 2012 – Fortschritte und Herausforderungen in Deutschland*. Münster: Waxmann.

Prosch, A. (in Vorb.). Entwicklung von Rechtschreibkompetenz. Differentielle Analysen mit NEPS-Daten der Haupterhebungen in Klasse 5 und 7 sowie den Entwicklungsstudien in Klasse 6 und 7. Universität Hamburg.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), Proceeds of the Forth Berkley Symposium on Mathematical Statistics and Probability (Vol.4, pp. 321–333). Berkley: University of California Press.

Rasch, B., Friese, M., Hofmann, W. & Naumann, E. (2006). *Quantitative Methoden 1. Einführung in die Statistik* (2. Aufl.). Heidelberg: Springer.

Rauch, D. P. & Hartig, J. (2011). Interpretation von Testwerten in der IRT. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 240–250). Heidelberg: Springer.

Reinders H. & Ditton, H. (2011). Überblick Forschungsmethoden. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz. (Hrsg.), *Empirische Bildungsforschung. Strukturen und Methoden* (Bd. 1, S. 45–51). Wiesbaden: VS Verlag für Sozialwissenschaften.

Reinders, H., Ditton, H., Gräsel, C. & Gniewosz, B. (2011a). *Empirische Bildungsforschung. Strukturen und Methoden* (Bd. 1). Wiesbaden: VS Verlag für Sozialwissenschaften.

Reinders, H., Ditton, H., Gräsel, C. & Gniewosz, B. (2011b). *Empirische Bildungsforschung. Gegenstandsbereiche* (Bd. 2). Wiesbaden: VS Verlag für Sozialwissenschaften.

Rieder, O. (1992). *Rechtschreibtest für 6. und 7. Klassen : RST 6–7* (2. Aufl.). Weinheim: Beltz.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 42–106). Weinheim: Beltz.

Rolff, H. G. (1993). *Wandel durch Selbstorganisation*. Weinheim: Juventa.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2. Aufl.). Bern: Huber.

Roth, H. & Aebli, H. (1969). *Begabung und Lernen. Ergebnisse und Folgerungen neuer Forschungen* (Bd. 4, 3. Aufl.). Stuttgart: Klett.

Sälzer, C., Reiss, K., Schiepe-Tiska, A., Prenzel, M. & Heinze, A. (2013). Zwischen Grundlagenwissen und Anwendungsbezug: Mathematische Kompetenz im internationalen Vergleich. In M. Prenzel, C. Sälzer, E. Klieme & O. Köller (Hrsg.), *PISA 2012 – Fortschritte und Herausforderungen in Deutschland* (S. 47–98). Münster: Waxmann.

Scherer, R. (2012). Analyse der Struktur, Messinvarianz und Ausprägung komplexer Problemlösekompetenz im Fach Chemie: Eine Querschnittstudie in der Sekundarstufe I und am Übergang zur Sekundarstufe II. Berlin: Logos Verlag.

Schneider, W. (Hrsg.). (2008). *Entwicklung von der frühen Kindheit bis zum frühen Erwachsenenalter*. Weinheim: Beltz.

Schneider, W., Marx, H. & Hasselhorn, M. (2008). *Diagnostik von Rechtschreibleistungen und -kompetenz* (Bd. 6). Göttingen: Hogrefe.

Schneider, W., Stefanek, J. & Dotzler, H. (1997). Erwerb des Lesens und Rechtschreibens: Ergebnisse aus dem SCHOLASTIK-Projekt. In F. E. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S. 113–130). Weinheim: Beltz.

Schnell, R., Hill, P. B. & Esser, E. (2011). *Methoden der empirischen Sozialforschung* (9. Aufl.). München: Oldenburg.

Schönweiß, F. (2004). *Münsteraner Rechtschreibanalyse*. Münster.

Schott, F. & Azizi Ghanbari, S. (2008). *Kompetenzdiagnostik, Kompetenzmodelle, kompetenzorientierter Unterricht. Zur Theorie und Praxis überprüfbarer Bildungsstandards. ComTrans – ein theoriegeleiteter Ansatz zum Kompetenztransfer als Diskussionsvorlage*. Münster: Waxmann.

Schweizer, K. (2006). *Leistung und Leistungsdiagnostik*. Berlin: Springer Medizin Verlag.

Seastrom, M. M. (2003). *NCES statistical standards*. Washington, D.C: National Center for Education Statistics. Retrieved May 15, 2014, from http://nces.ed.gov/pubs2003/2003601.pdf

Seidel, T. & Prenzel, M. (2008). Assessment in large scale studies. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of competencies in educational contexts. State of the art and future prospects* (S. 279–304). Göttingen: Hogrefe & Huber.

Stock, C. & Schneider, W. (2008). *DERET 3/4+: Deutscher Rechtschreibtest für das dritte und vierte Schuljahr*. Göttingen: Hogrefe Verlag.

Stubbe, T. C., Tarelli, I. & Wendt, H. (2012). Soziale Disparitäten der Schülerleistungen in Mathematik und Naturwissenschaften. In W. Bos, H. Wendt, O. Köller & C. Selter (Hrsg.), *TIMSS 2011 – Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 231–246). Münster: Waxmann.

Tarelli, I., Lankes, E.-M., Drossel, K. & Gegenfurtner, A. (2012). Lehr- und Lernbedingungen an Grundschulen im internationalen Vergleich. In W. Bos, I. Tarelli, A. Bremerich-Vos & K. Schwippert (Hrsg.), *IGLU 2011 – Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 137–173). Münster: Waxmann.

Tarelli, I., Schwippert, K. & Stubbe, T. C. (2012). Mathematische und naturwissenschaftliche Kompetenzen von Schülerinnen und Schülern mit Migrationshintergrund. In W. Bos, H. Wendt, O. Köller & C. Selter (Hrsg.), *TIMSS 2011 – Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 247–267). Münster: Waxmann.

Thomé, G. & Gomolka, J. (2007). Rechtschreiben. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistung International)* (S. 140–146). Weinheim: Beltz.

Thomé, G. & Thomé, D. (2010). *OLFA 3–9: Oldenburger Fehleranalyse für die Klassen 3–9. Instrument und Handbuch zur Ermittlung der orthographischen Kompetenz aus freien Texten und für die Planung und Qualitätssicherung von Fördermaßnahmen*. Oldenburg: isb.

Thurgood, L. Walter, E., Carter, G., Henn, S., Huang, G., Nooter, D., Smith, W., Cash, R. W. & Salvucci, S. (2003). *NCES handbook of survey methods (Technical report)*. Washington, DC: National Center for Education Statistics. Retrieved May 15, 2014, from http://nces.ed.gov/pubs2003/2003603.pdf

Tippelt, R. & Schmidt, B. (2010). *Handbuch Bildungsforschung* (3. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Valtin, R., Badel, I., Löffler, I., Meyer-Schepers, U. & Voss, A. (2003). Orthographie als Lerngegenstand. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, G. Walther & R. Valtin (Hrsg.), *Erste Ergebnisse aus IGLU* (S. 227–264). Münster: Waxmann.

Vieluf, U., Ivanov, S. & Nikolova, R. (2011). *KESS 10/11 – Kompetenzen und Einstellungen von Schülerinnen und Schüler an Hamburger Schulen am Ende der Sekundarstufe I und zu Beginn der gymnasialen Oberstufe*. Münster: Waxmann.

von Davier, A. A., Carstensen, C. H., & von Davier, M. (2008). Linking Competencies in Horizontal, Vertical and Longitudinal Settings and Measuring Growth. In J. Hartig, E. Klieme, & D. Leutner (Hrsg.), *Assessment of Competencies in Educational Contexts* (S. 53–80). New York: Hogrefe & Huber.

Voss, A. (2006). *Print- und Hypertextlesekompetenz im Vergleich. Eine Untersuchung von Leistungsdaten aus der Internationalen Grundschul-Lese-Untersuchung (IGLU) und der Ergänzungsstudie Lesen am Computer (LaC).* Münster: Waxmann.

Voss, A. (2009). Zur Erfassung und Modellierung von Rechtschreibkompetenz. In R. Valtin & B. Hofmann (Hrsg.), *Kompetenzmodelle der Orthographie. Empirische Befunde und förderdiagnostische Möglichkeiten* (S. 12–25). Berlin: DGLS.

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, 13* (23), 15–33.

Voss, A., Carstensen, C. H. & Bos, W. (2005). Textgattungen und Verstehensaspekte. Analyse von Leseverständnis aus den Daten der IGLU-Studie. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin & G. Walther (Hrsg.), *IGLU – Vertiefende Analysen zu Leseverständnis, Rahmenbedingungen und Zusatzstudien* (S. 1–36). Münster: Waxmann.

Voss, A., Löffler, I., Meyer-Schepers, U., Kowalski, K. & Meckel, C. (2008). Frühdiagnose rechtschreibschwächerer Schülerinnen und Schüler auf der Grundlage von Kompetenzmodellen. In W. Bos, H. G. Holtappels, H. Pfeiffer, H.-G. Rolff & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung, Band 15. Daten, Beispiele und Perspektiven* (S. 123–155). Weinheim und München: Juventa.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54* (3), 427–450.

Weinert, F. E. (1999). *Concepts of Competence. Definition and selection of competencies*. Paris: OECD.

Weinert, F. E. (2001). *Leistungsmessungen in Schulen*. Weinheim: Beltz-Verlag.

Weinert, F. E. & Helmke, A. (1997). *Entwicklung im Grundschulalter*. Weinheim: Psychologie Verlagsunion.

Wendt, H., Stubbe, T. C. & Schwippert, K. (2012). Soziale Herkunft und Lesekompetenzen von Schülerinnen und Schülern. In W. Bos, I. Tarelli, A. Bremerich-Vos & K. Schwippert (Hrsg.), *IGLU 2011 – Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 175–190). Münster: Waxmann.

Wilson, M. (2005). *Construction Measures: An Item Response Modeling Approach*. Mahwah, New Jersey, London: Lawrence Erlbaum Associates.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). *ACER ConQuest Version 2.0 Generalised item response modelling software*. Melbourne: Acer Press.

Zedler, P. & Döbert, H. (2010). Erziehungswissenschaftliche Bildungsforschung. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (S. 23–45). Wiesbaden : VS Verlag für Sozialwissenschaften.

Zöfel, P. (2002). *Statistik verstehen. Ein Begleitbuch zur computer-gestützten Anwendung*. München: Addison-Wesley Verlag.

11. ANHANG

Abbildung 11.1: Verteilung der Personenfähigkeit für das ganze Wort der ES K7 (N = 307)

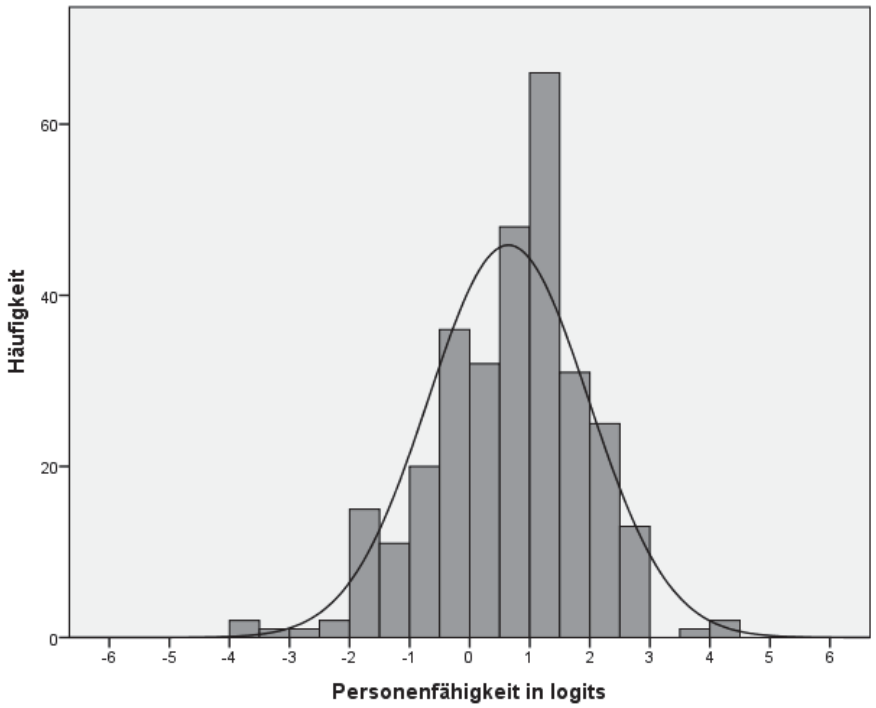


Abbildung 11.2: Grafischer Modelltest auf Basis der Itemschwierigkeiten – ES K7

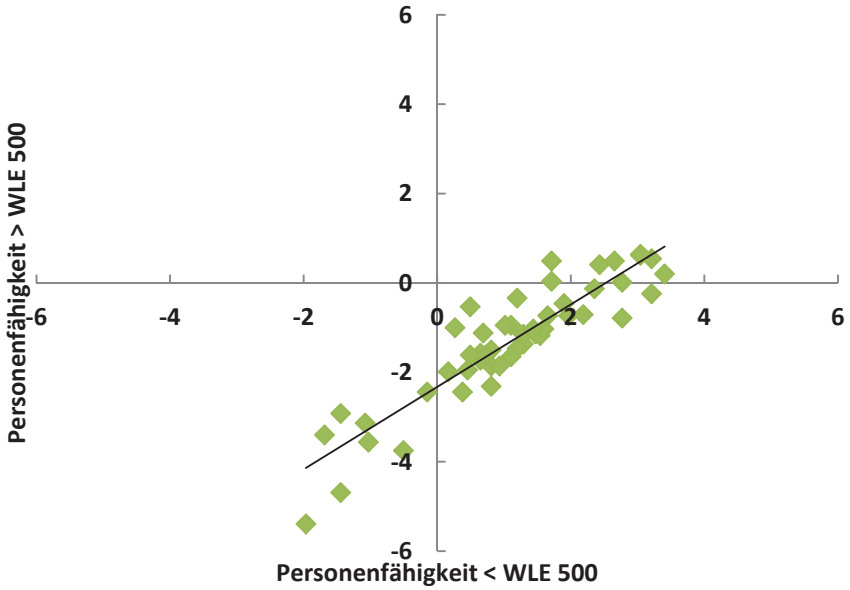
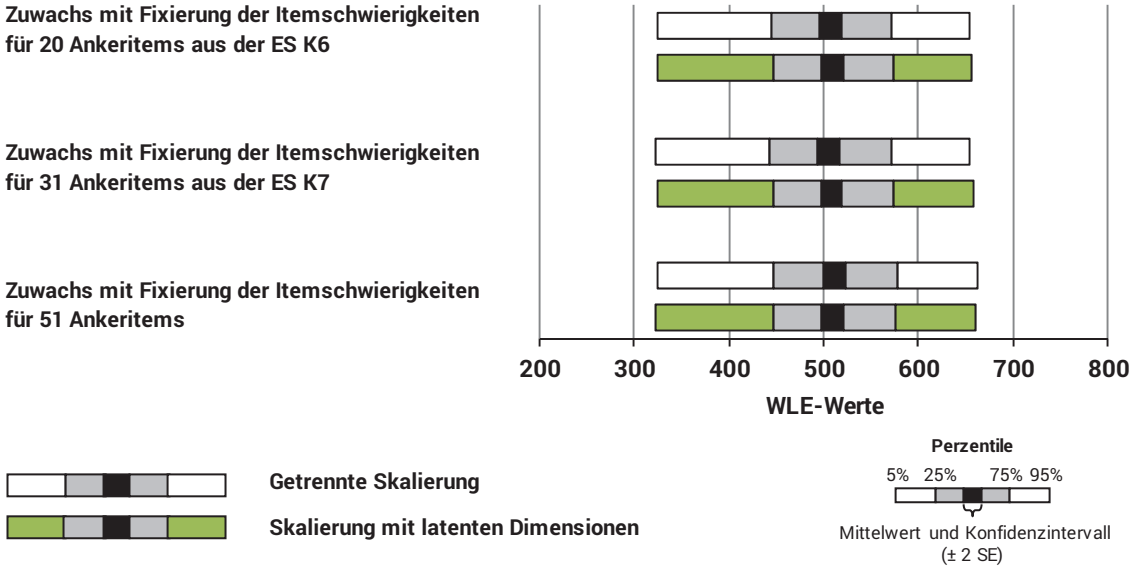


Abbildung 11.3: Vergleich der Skalierungsverfahren mit zusätzlichen zeitpunktspezifischen Items



ZUSAMMENFASSUNG

In der methodisch ausgerichteten Arbeit mit dem Titel „Quantitative Datenanalyse zur längsschnittlichen Erfassung der Rechtschreibkompetenz in NEPS unter besonderer Berücksichtigung der Kompetenzstruktur und der Einflussfaktoren“ wird aus einer erziehungswissenschaftlichen Perspektive grundlegenden Fragen zur längsschnittlichen Kompetenzmessung und -modellierung der Rechtschreibung nachgegangen. Dazu werden Daten aus dem Nationalen Bildungspanel (National Educational Panel Study, NEPS) genutzt, die in den Entwicklungsstudien in den Klassenstufen 5 bis 7 und der Haupterhebung in Klassenstufe 5 erhoben wurden.

Die Zielsetzung der Arbeit ist theoretisch und empirisch im aktuellen Forschungskontext verankert. Seit der Teilnahme Deutschlands an international vergleichenden Leistungsstudien, die querschnittlich angelegt sind, hat die empirische Bildungsforschung an Bedeutung gewonnen, und es etablieren sich in der empirischen Bildungsforschung auch Bildungspanel bzw. Längsschnittstudien. Im Zusammenhang mit der „empirischen Wende“ vollzog sich eine Abwendung von der reinen Leistungsmessung hin zu einer kompetenzorientierten Leistungsmessung. Dies setzt theoretisch begründete Kompetenzmodelle und eine empirische Validierung mit psychometrischen Verfahren voraus. Dafür sind geeignete Messkonzepte und -verfahren notwendig.

Die Kompetenzmessung der Rechtschreibung im Rahmen der Längsschnittstudie NEPS erfolgt mit einem sprachsystematischen Rechtschreibtest (SRT). Er basiert auf einem differenziellen sprachsystematischen Rechtschreibkompetenzmodell, das theoretisch auf Ergebnissen der linguistischen Graphematik aufbaut und mit Daten aus Querschnittstudien empirisch überprüft wurde. Der SRT wird für den Einsatz in der NEPS-Studie zur Kompetenzmessung in der Sekundarstufe I weiterentwickelt, um geeignete längsschnittliche Testinstrumente für die kompetenzorientierte Leistungsmessung zu entwickeln und adäquate Verfahren für die Modellierung der Rechtschreibkompetenz zu finden. Dabei gilt es grundsätzliche Fragen zu klären, denen sich die vorliegende Arbeit hinsichtlich ihrer Forschungsziele widmet:

- Haben die in einer Entwicklungsstudie nachgewiesenen Gütekriterien für das Messinstrument SRT auch in der folgenden repräsentativen Haupterhebung Bestand?

- Welche Analyseverfahren eignen sich für die längsschnittliche Kompetenzentwicklung, um diese verlässlich und effizient zu erfassen und mögliche Änderungen der differenziellen Kompetenzstruktur aufzudecken?
- Welche Faktoren wirken sich auf die differenzielle Kompetenzentwicklung aus?

Zur Beantwortung der Fragen werden im Sinne der interdisziplinären Forschung empirische und didaktische Erkenntnisse für die quantitative Datenanalyse zur längsschnittlichen Erfassung der Rechtschreibkompetenz kombiniert.

Hinsichtlich der Testentwicklung wurde aufgezeigt, dass die verhältnismäßig kleine Entwicklungsstudie zur Erprobung des sprachsystematischen Rechtschreibtests zu verlässlichen Ergebnissen führte, da sich die ermittelten Gütekriterien nur unwesentlich von den in der Haupterhebung ermittelten unterschieden. Der in dieser Arbeit angestellte Vergleich der Verfahren zur längsschnittlichen Erfassung der Rechtschreibkompetenz liefert vertiefende Erkenntnisse im Hinblick auf ihre Eignung, wobei unter Abwägung von Vor- und Nachteilen eine Kombination zweier Verfahren besonders verlässlich und effizient erscheint. Die Untersuchung zur Entwicklung der Kompetenzstruktur der Rechtschreibung von Klassenstufe 6 bis 7 liefert Hinweise für unterschiedliche differenzielle Kompetenzentwicklungen der Schülerinnen und Schüler. Dabei zeigte sich, dass die Entwicklung der Rechtschreibkompetenz in der Sekundarstufe I insbesondere vom Kompetenzstand in den Teilkompetenzen und weniger von Einflussfaktoren abhängt. Auf diesen Ergebnissen kann eine weiterführende Forschung mit den Daten der Haupterhebungen über den gesamten Zeitraum der Sekundarstufe I aufbauen, deren Ergebnisse wiederum die rechtschreibdidaktische Forschung entscheidend voranbringen können. Die gewonnenen methodischen Erkenntnisse können dazu beitragen, die quantitative Datenanalyse zur längsschnittlichen Erfassung der Rechtschreibkompetenz beeinflussen, um sie in zukünftigen Studien erneut auf ihre Evidenz hin zu überprüfen.

EIDESSTATTLICHE VERSICHERUNG

Ich versichere hiermit, dass ich die vorliegende Arbeit mit dem Titel „Quantitative Datenanalyse zur längsschnittlichen Erfassung der Rechtschreibkompetenz in NEPS unter besonderer Berücksichtigung der Kompetenzstruktur und der Einflussfaktoren“ ohne unzulässige Hilfe Dritter und ohne Nutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Sämtliche in dieser Arbeit wörtlich oder inhaltlich übernommenen Stellen sind als solche kenntlich gemacht.

Die Arbeit oder Teile davon wurden bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Ferner versichere ich, mich bisher keiner Doktorprüfung an einer Hochschule unterzogen oder um Zulassung zu einer solchen beworben zu haben.

Dortmund, den

_____