# Die Validierung von Rechtschreibkompetenzmodellen im Rahmen einer empirisch vergleichenden Analyse orthografischer Leistungstests

**Dissertation**

zur Erlangung des Grades eines

D o k t o r s  d e r  P h i l o s o p h i e

der Technischen Universität Dortmund
an der Fakultät Erziehungswissenschaft,
Psychologie und Soziologie

vorgelegt von
Kerstin Naujokat

Dortmund, im Februar 2015

Gutachter:
Prof. Dr. Wilfried Bos
Prof. Dr. Albert Bremerich-Vos

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# TABELLENVERZEICHNIS

# 1 Einleitung und Überblick

Rechtschreibung begeistert und entgeistert. Betrachtet man Kinder im vorschulischen Bereich, so lässt sich die Freude und Begeisterung erkennen, mit denen sie z. B. ihren eigenen Namen das erste Mal schreiben. Stolz zeigen sie die Schriftprodukte ihren Eltern, Verwandten und Freunden. Von der Schule erhalten Schülerinnen und Schüler zunächst für die Verschriftung einzelner Buchstaben, schließlich für ganze Wörter und letztendlich für fehlerfreie oder mit wenigen Fehlern behaftete Schreibungen gute Bewertungen, Benotungen und Lob.

Leider trifft dies nicht auf alle Schriftlernenden zu. Groß angelegte Schulleistungsstudien wie IGLU, KESS und die IQB-Ländervergleiche[1] haben ernüchternde und unbefriedigende Ergebnisse zum Leistungsstand im Bereich Rechtschreibung ans Licht gebracht. Ein nicht unerheblicher Teil von Kindern verfügt nicht über die Fertigkeiten und Fähigkeiten, die auf institutioneller Ebene vorausgesetzt werden. Ferner zeigt sich eine große Leistungsschere zwischen rechtschreibstarken und rechtschreibschwachen Schülerinnen und Schülern. Diese Beobachtungen beschränken sich dabei nicht nur auf den Primarbereich, sondern weiten sich auch auf die Sekundarstufe aus (DESI[2], IQB-Ländervergleich in Jahrgangsstufe 9). In der Öffentlichkeit haben Ergebnisse zur Leistungsentwicklung von Kindern und Jugendlichen ebenfalls für Aufsehen gesorgt. So zeigt beispielsweise eine über den Zeitraum von 40 Jahren geführte Studie (Steinig & Betzel, 2014, S. 362; Steinig, Betzel, Geider & Herbold, 2009, S. 252 ff.), dass sich die Leistungen im generationsübergreifenden Vergleich im Durchschnitt verschlechtert haben (die Längsschnittstudie LOGIK[3] weist diese Entwicklung ebenfalls aus (Schneider & Stefanek, 2007, S. 80)). Dabei ist

[1] IGLU wird in Abschnitt 1.1 sowie KESS und der IQB-Ländervergleich in Abschnitt 2.1 dargestellt.

[2] DESI steht für *Deutsch-Englisch-Schülerleistungen-International*. Jugendliche aus der neunten Jahrgangsstufe wurden 2003/2004 bundesweit in den Fächern Deutsch und Englisch getestet. Insgesamt haben etwa 11.000 Schülerinnen und Schüler an der Studie teilgenommen (Beck, Bundt & Gomolka, 2008, S. 11).

[3] LOGIK ist die Münchener *Longitudinalstudie zur Genese individueller Kompetenzen*. Die Studie begleitete zunächst über neun Jahre hinweg und schließlich in Nachuntersuchungen die Entwicklungen u. a. der Schriftsprachkompetenz von Kindern bis ins junge Erwachsenenalter. Zu Beginn der Erhebung im Jahr

Rechtschreibung eine Kulturtechnik, deren Vermittlung Auftrag der Schule ist und in dieser Lernumwelt sichergestellt werden sollte, um die Schülerinnen und Schüler auf zukünftige Anforderungen in Ausbildung und Beruf vorzubereiten. In den KMK-Bildungsstandards für das Fach Deutsch in der Primarstufe wird der Bereich „Schreiben“ als separate Kategorie aufgeführt, worunter auch „richtig schreiben“ fällt (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2005b, S.7).

Die Verankerung in den Bildungsstandards ist in Anbetracht der Wichtigkeit des Erlernens der Rechtschreibung für den privaten, schulischen und beruflichen Erfolg selbstredend. So belegen Untersuchungen (wie z. B. IGLU, KESS, LOGIK, DESI) den direkten und indirekten Einfluss des orthografisch korrekten Schreibens sowohl auf andere Fächer als auch auf die Schulformwahl und bestätigen damit die große Auslesebedeutung, die bereits vor über 40 Jahren von Kemmler (1970, S. 126 ff.) konstatiert wurde und hiermit immer noch aktuell ist (Schneider, 2008b, S. 147; P. Marx, 2007, S. 114 f.; Schneider & Stefanek, 2007, S. 80; Valtin, Badel, Löffler, Meyer-Schepers & Voss, 2003, S. 248; May, 2002a, S. 105). Rechtschreibung besitzt eine hohe prädiktive Kraft für die schulische Laufbahn, die sogar die der Leseleitung übersteigt, und hat damit einen großen Stellenwert (Kowalski, Voss, Valtin & Bos, 2010, S. 33; Beck, Thomé & Thomé, 2009, S. 45).

Trotz dieser Wichtigkeit der Rechtschreibung ist sie einer der unbeliebtesten Lernbereiche und häufig negativ besetzt, indem sie als irregulär betrachtet, von Rechtschreibchaos oder -katastrophe gesprochen und zeitweise ihr systematischer Charakter überhaupt in Frage gestellt wird (Munske, 2005, S. 9, 18 f.; Valtin, Badel et al., 2003, S. 227; Hinney, 1997, S. 58 ff.). Als Erklärung weisen Schneider, Marx und Hasselhorn auf das Verhältnis von Graphem-Phonem-Korrespondenzen und Phonem-Graphem-Korrespondenzen hin. So schreiben sie, dass die Zahl der Graphem-Phonem-Korrespondenzen relativ überschaubar ist, also ein Graphem durch eine begrenzte Anzahl an Lauten wiedergegeben wird, während die Anzahl der Phonem-Graphem-Korrespondenzen im Gegensatz als groß bezeichnet wird, d. h. ein wahrgenommener Laut entspricht vielen regelkonformen Graphemen. Sie verdeutlichen dies an den theoretisch unterschiedlichen möglichen Schreibversionen des Wortes Fuchs: *Vuks, *Fux, *Vuchs etc. (Hasselhorn, Marx & Schneider, 2008, S. 1). Die in manchen Köpfen noch verankerte (Wunsch-)Vorstellung einer 1:1-Abbildbarkeit von Phonem und Graphem, also einer Alphabetschrift mit der dahinterliegenden Maxime „Schreib, wie Du sprichst.“, ist sicherlich auch ein Grund für den Unmut gegenüber der Rechtschreibung.

Im Rahmen von Leistungsstudien zur Schriftsprache wird die Orthografie im Vergleich zum Lesen eher stiefmütterlich behandelt. Allein ein Blick auf die Kapitelüberschriften der Berichtsbände von Large-Scale-Assessments zeigt, dass die Rechtschreibung zurückhaltender und zum Teil auch gar nicht betrachtet wird. Schneider (2008b, S. 145) spricht sogar von Orthografie als „second class skill“. Dabei erzeugt die Orthografie heftige Diskussionen innerhalb des Forschungsfelds. Kontroverse Standpunkte zu Merkmalen und zum Erwerb werden vertreten sowie unterschiedliche Modelle formuliert (Hinney, 1997,

1984 setzte sich die Stichprobe aus 200 Kindern im Alter von vier Jahren zusammen (Schneider, 2008a, S. 171 ff.; Schneider, 2008b, S. 147).

S. 57 ff.). Es fehlt eine verbindliche und konforme Vorstellung davon, wie das Konstrukt der orthografischen Kompetenz zu konzeptualisieren und quantifizieren ist.

Letztendlich streben alle Beteiligten das übergeordnete Ziel an, dass Kinder erfolgreich lernen, richtig zu schreiben. Dabei geraten manche Eltern und Lehrpersonen in eine Notlage. Trotz vielfältiger und intensiver Bemühungen scheitern ihre Kinder am Lerngegenstand Rechtschreibung. So investieren beispielsweise in einer Schulwoche die Hälfte aller Viertklässler mindestens zwei reine Zeitstunden in den Rechtschreibunterricht, ein Drittel sogar mehr als drei (Valtin, Löffler, Meyer-Schepers & Badel, 2004a, S. 152 ff.; Valtin, Badel et al., 2003, S. 242). Daher überrascht es auch nicht, dass auf (vermeintliche) Hilfen zurückgegriffen wird und die Nachfrage nach sowie die Produktion von – z. T. kommerziellen – diagnostischen Verfahren zur Erfassung von Rechtschreibleistung und Förderansätzen sowie -materialien boomt (Hasselhorn et al., 2008, S. 3).

In diesem Zusammenhang wird die Wichtigkeit theoretisch durchdachter und empirisch überprüfter Kompetenzmodelle besonders greifbar. Dabei ist die Zusammenarbeit von Fachdidaktik und empirischer Bildungsforschung für didaktisch fundierte und empirisch validierte Aussagen unerlässlich. Validierte Kompetenzmodelle zur Rechtschreibung sind in der Gegenwart jedoch keine Selbstverständlichkeit, obwohl die Modellierung und Operationalisierung von Kompetenz entscheidende Auswirkungen darauf haben kann, wie gelehrt wird. Wenn beispielsweise der Rechtschreiberwerb in aufeinanderfolgenden Stufen verläuft, kann es einen hinderlichen bis schädlichen Einfluss haben, morphologische Ableitungsstrategien zu früh zu erläutern. Falls sich beim Erwerb hingegen alle Zugriffsweisen auf die Schrift zusammen entwickeln, wäre es nicht sinnvoll, die rechtschriftlichen Stoffgebiete strikt hintereinander abzuarbeiten. Aus den Hypothesen zum Erwerb werden Lern- und Lehrwege formuliert, die Einfluss auf die Sicherheit der Schreibung der Schriftlernenden haben. Um wirksame Lernhilfen und Förderungen anzusetzen, ist es deshalb wichtig, dass die Hypothesen bzw. Modelle zur Rechtschreibung überprüft werden.

## 1.1 Einordung und Zielsetzung der ZuRecht-Studie

Die vorliegende Arbeit ist im Rahmen der *Internationalen Grundschul-Lese-Untersuchung* (IGLU) entstanden. IGLU ist eine Schulleistungsstudie, die das Leseverständnis von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in einer für Deutschland repräsentativen Stichprobe erforscht. Seit ihrem Auftakt 2001 wird die Studie im Abstand von fünf Jahren regelmäßig wiederholt. Auf internationaler Ebene wird IGLU von der *Association for the Evaluation of Educational Achievement* (IEA) unter der Bezeichnung *Progress in International Reading Literacy Study* (PIRLS) durchgeführt (Bos, Hornberg et al., 2008, S. 9).

IGLU wurde um die nationale Ergänzungsstudie IGLU-E erweitert, um zusätzliche Erkenntnisse – unter anderem zu den Kompetenzen von Schülerinnen und Schülern im Bereich Orthografie – zu gewinnen. Die wissenschaftliche Leitung von IGLU und IGLU-E in Deutschland hat Wilfried Bos vom Institut für Schulentwicklungsforschung (IFS) der

Technischen Universität Dortmund. Aufgrund der ernüchternden Ergebnisse von IGLU-E 2001 wurde der Rechtschreibung im darauffolgenden Zyklus eine verstärkte Aufmerksamkeit gewidmet. Die vorliegende Arbeit stellt eines der Resultate dar. Initiiert wurde eine Zusatzstudie zu IGLU-E für den Bereich der Rechtschreibung[4]. Sie trägt das Akronym „ZuRecht“ (**Zu**satzstudie **Recht**schreibung).

Inhaltlich und methodisch knüpft die Arbeit an die Orthografieuntersuchung in IGLU-E 2001 an und erweitert die Ergebnisse zu den Schülerleistungen um das Ziel der Validierung und des empirischen Vergleichs zweier Rechtschreibtests. Bei den beiden schriftsprachlichen Tests handelt es sich um den *gutschrift-diagnose-Test*, der u. a. bereits in der ersten IGLU-Erhebung (IGLU-E 2001) eingesetzt wurde, sowie um den *Sprachsystematischen Rechtschreibtest* (SRT), der u. a. in der IGLU-Voruntersuchung 2006 und im *Nationalen Bildungspanel*[5] (NEPS) verwendet wurde bzw. wird.

Die oben genannte übergeordnete Zielsetzung gliedert sich in mehrere Bestrebungen. Zunächst sollen die den beiden Tests zugrunde liegenden Kompetenzmodelle[6] mit den aktuellen statistischen Auswertungsverfahren, wie sie auch in Large-Scale-Assessments zur Modellierung von kognitiven Leistungsdaten zum Einsatz kommen, empirisch überprüft werden. Hierbei ist zu beachten, dass der SRT u. a. von der Autorin bereits in der Voruntersuchung zu IGLU 2006 pilotiert wurde (Kowalski, 2007; Voss, Blatt & Kowalski, 2007). Daher soll auch geprüft werden, ob die Ergebnisse reproduziert werden können. Erweitert werden die Analysen des SRT durch die Einbindung der IGLU-Hauptuntersuchung 2006. Über diesen deutlich größeren Referenzrahmen soll ferner eine Kontrolle der Befunde dieser kleineren Untersuchung erfolgen. Im Fokus der Arbeit steht die vergleichende Auseinandersetzung mit den beiden Rechtschreibtests. Dabei werden psychometrische Kriterien der Testkonstruktion und Modellstruktur analysiert und gegenübergestellt, um Aussagen über die Güte der Tests formulieren zu können. Auswertungen des rechtschriftlichen Kompetenzstandes der Schülerinnen und Schüler, des Zusammenhangs mit ausgewählten Hintergrundvariablen sowie die Identifizierung spezifischer Profilverläufe ergänzen die statistischen Prüfverfahren. Hierbei sollen stets die differenziellen Teilkompetenzen der Rechtschreibmodelle berücksichtigt werden. Das heißt, dass die Schreibungen klassifiziert und qualitativ analysiert werden, was bei dem Großteil der orthografischen Studien bisher nicht oder nur ansatzweise durchgeführt wurde, um auch auf diesem Weg Aufschlüsse über die latente Rechtschreibkompetenzstruktur zu erhalten.

[4] Hiermit ist die kodifizierte Form des Schreibens gemeint; Aspekte des schriftlichen Ausdrucks sind nicht Gegenstand dieser Arbeit.

[5] NEPS bezeichnet die *National Educational Panel Study*. Es handelt sich hierbei um eine Längsschnittstudie, die die Bildungsprozesse und Kompetenzentwicklungen vom frühen Kindes- bis ins späte Erwachsenenalter untersucht. Seit 2009 werden Ziel- und Kontextpersonen aus ganz Deutschland befragt (Blossfeld, von Maurice & Schneider, 2011, S. 6 ff.).

[6] Rechtschreibkompetenz wird in dieser Arbeit als latentes Konstrukt betrachtet, das nicht direkt beobachtbar ist, sondern über manifeste Variablen erschlossen werden muss (vgl. Abschnitt 3.3.1). Die manifesten Variablen, die dafür genutzt werden, stammen aus der Datenerhebung mittels Rechtschreibtests. Es sind die beobachteten Rechtschreibleistungen der Schülerinnen und Schüler, die über Kompetenzmodelle beschrieben werden. Dementsprechend wird der Begriff der Kompetenz verwendet, wenn es um die Analyse und Ergebnisse der Daten mittels Kompetenzmodellen geht.

## 1.2 Forschungsfragen und Methodik

Um die genannten Zielsetzungen aus Abschnitt 1.1 zu konkretisieren, erfolgt aus diesen eine Ableitung von Forschungsfragen, die im Rahmen der vorliegenden Dissertation beantwortet werden. Sie unterteilen sich jeweils in eine übergeordnete Forschungsfrage, aus der sich mehrere Subfragen ergeben. Erweitert werden die Forschungsfragen über die Angabe der zur Beantwortung verwendeten Auswertungsmethodik.

Methodologische Grundlage für die Analyse und den Vergleich der aus den Rechtschreibtests gewonnenen Daten bilden *probabilistische Testverfahren* (Rost, 2004; Embretson & Reise, 2000; Fischer, 1974; G. Rasch, 1960). Die probabilistische Testtheorie bietet die Möglichkeit, die linguistischen Annahmen zum Aufbau orthografischer Leistung in Form eines Kompetenzmodells auszudifferenzieren, Einsichten in die spezifischen Eigenschaften der eingesetzten Orthografietests zu gewinnen und die Datenanpassung an die theoretischen Konzepte zu ermitteln. Darüber hinaus erlauben probabilistische Testverfahren, Itemschwierigkeiten (Schwierigkeit der orthografischen Analyseeinheiten) und Personenfähigkeiten (orthografische Kompetenz der Kinder) auf einer gemeinsamen Skala abzubilden, um diese Parameter direkt aufeinander zu beziehen, was für Analysen des Zusammenhangs mit Hintergrundmerkmalen vorteilhaft ist.

Im Folgenden sind die Forschungsfragen aufgeführt. Dabei sind jeder Subfrage auch einzelne Analyseschritte zugeordnet. Ergänzt werden die Forschungsfragen durch Verweise auf das entsprechende Kapitel bzw. den entsprechenden Abschnitt, in denen sie behandelt werden.

**I Welche Rechtschreibkompetenzen weisen die im Rahmen von ZuRecht getesteten Grundschülerinnen und Grundschüler auf?**

Auswertungsmethodik: Deskriptive Statistiken auf Wort- und Teilkompetenzebene

| Subfrage | Analyseschritt | Abschnitt |
|---|---|---|
| a) Welche Leistungen erbringen die Schülerinnen und Schüler auf Ganzwortebene und in den Teilkompetenzen? | Häufigkeitsauszählungen richtig geschriebener Wörter sowie der Analyseeinheiten der Subskalen | 4.1, 4.2.3 |
| b) Die Schreibung welcher Wörter fällt den Kindern besonders leicht bzw. schwer? | Auszählung der Variantenschreibungen sowie Betrachtung häufiger Fehlerquellen | 4.1 |

**II Können die theoriekonformen mehrdimensionalen Rechtschreibkompetenzmodelle empirisch belegt werden?**

Auswertungsmethodik: Skalierung der Daten mit einparametrigen logistischen IRT-Modellen: Vergleich ein- und mehrdimensionaler dichotomer und ordinaler Raschmodelle

| Subfrage | Analyseschritt | Abschnitt |
|---|---|---|
| a) Sind die Items modellkonform? | Betrachtung der Itemfit-Werte | 4.2.1, 4.2.2 |
| b) Sind die Tests bezüglich ihres Schwierigkeitsgrades angemessen? | Vergleich von Personenfähigkeiten und Itemschwierigkeiten | 4.2.1, 4.2.2 |
| c) Können die mehrdimensionalen Strukturen der Kompetenzmodelle empirisch belegt werden? | Vergleich der Datenanpassung konkurrierender Modellvarianten anhand informationstheoretischer Maße sowie globaler Modellgeltungstests (Likelihoodquotiententest und $\chi^2$-verteilte Prüfstatistik) | 4.2.1, 4.2.2 |
| d) Sind die der Theorie nach ausgewiesenen Teilkompetenzen eigenständig? | Analyse der Korrelationsmatrizen sowie Skalierung mit arbiträrer Itemklassifikation | 4.2.1, 4.2.2 |

**III Welche konzeptionellen und psychometrischen Gemeinsamkeiten und Unterschiede weisen die Rechtschreibtests bei einem Vergleich auf?**

Auswertungsmethodik: quantitative Inhaltsanalyse, testübergreifende Skalierung

| Subfrage | Analyseschritt | Abschnitt |
|---|---|---|
| a) Welche Rechtschreibphänomene werden welchen Teilkompetenzen der Tests zugeordnet? | inhaltsanalytische Zuordnung und Auszählung der Indikatoren zu den Teilkompetenzen sowie theoretischer Vergleich der Konzepte | 2.4.1, 2.4.2 |
| b) Wie hoch ist der Zusammenhang zwischen den Teilkompetenzen der Tests untereinander? | Berechnung eines neundimensionalen IRT-Modells zur Analyse der latenten Interkorrelationen zwischen den Teilkompetenzen der Tests | 4.2.3 |
| c) Wie zuverlässig lassen sich die definierten Teilkompetenzen im Vergleich messen? | Analyse der Reliabilitäten | 4.2.3 |

**IV Wie schneiden die Kinder im bundesweiten Vergleich ab und lassen sich die Ergebnisse aus ZuRecht zur Kompetenzmodellierung des SRT reproduzieren?**

Auswertungsmethodik: siehe Forschungsfragen I und II

| Subfrage | Analyseschritt | Abschnitt |
|---|---|---|
| a) Wie ist das orthografische Kompetenzniveau der Schülerinnen und Schüler aus ZuRecht im Verhältnis zum bundesweiten Kompetenzstand zu bewerten? | Verortung der Testleistungen auf Ganzwort- und Teilkompetenzebene aus ZuRecht in der IGLU-Haupterhebung über parallel durchgeführte Statistiken (vgl. I a)) | 4.2.4 |
| b) Subfragen II a) bis II d) entsprechend | Auswertung des SRT in der IGLU-Haupterhebung: Schritte aus Forschungsfrage II | 4.2.4 |

**V Welche der erhobenen Hintergrundmerkmale kennzeichnen gute Rechtschreiber und bestehen test- und teilkompetenzspezifische Unterschiede?**

Auswertungsmethodik: Berechnung der Parameterwerte der Personenfähigkeiten sowie statistische Beurteilung der Mittelwertsunterschiede

| Subfrage | Analyseschritt | Abschnitt |
|---|---|---|
| Welchen Zusammenhang hat die Rechtschreibkompetenz mit ... | Analyse und z-Standardisierung der Weighted-Likelihood-Estimates sowie Berechnung von Effektstärken nach Cohen | |
| a) dem Geschlecht? | | 4.3.1 |
| b) der weiterführenden Schulform? | | 4.3.2 |
| c) dem Bildungshintergrund? | | 4.3.3 |

**VI Können Schülergruppen identifiziert werden, deren Leistungsstand in den Teilkompetenzen der beiden Tests variiert und spezifische Verläufe aufweist?**

Auswertungsmethodik: explorative Einzelfallanalyse und probabilistische Clusteranalyse

| Subfrage | Analyseschritt | Abschnitt |
|---|---|---|
| a) Welche orthografischen Kompetenzen weisen die Tests den gleichen Kindern aus? | explorative Einzelfallanalyse und -interpretation | 4.4.1 |
| b) Können Gruppen von Kindern mit unterschiedlichen Kompetenzprofilen ermittelt werden? | latente Profilanalyse | 4.4.2 |

## 1.3 Aufbau der Arbeit

Die vorliegende Arbeit gliedert sich in einen theoretischen und einen empirischen Teil. In Kapitel 2 wird der aktuelle Forschungsstand zur differenziellen Erhebung orthografischer Kompetenz sowie gleichzeitiger empirischer Überprüfung der dafür verwendeten Rechtschreibtests ausführlich erläutert. Abschnitt 2.1 stellt Informationen zu dem Erhebungsrahmen dieser Studien, den Testkonzeptionen und den Ergebnissen zusammen. In den folgenden Abschnitten werden der gutschrift-diagnose-Test sowie der SRT beschrieben, die im Rahmen dieser Arbeit eingehend betrachtet werden. Der Aufbau von 2.2 und 2.3 gliedert sich in eine Darstellung der theoretischen Konzeption sowie des Testeinsatzes in bisherigen Erhebungen, womit die Schilderung des Forschungsstands aus 2.1 vervollständigt wird. Bei der Darstellung der linguistischen Konzepte werden die spezifischen Teilkompetenzen zur Modellierung von Rechtschreibkompetenz sowie die zur Operationalisierung verwendeten Wörter und Indikatoren beschrieben. Eine zusammenfassende Gegenüberstellung sowie kritische Diskussion von gutschrift-diagnose und dem SRT erfolgt gemeinsam mit einem Exkurs zu Ausnahmeschreibungen in Abschnitt 2.4 und schließt damit die schrifttheoretische Beschreibung der Orthografietests ab.

Eine Erläuterung des Erhebungsrahmens von ZuRecht eröffnet Kapitel 3. Hier werden u. a. Angaben zur Stichprobengröße dargelegt, die Diktattexte der beiden Rechtschreibtests vorgestellt und auf die Vorteile des Studiendesigns eingegangen (Abschnitt 3.1). Daran schließt sich eine Erklärung des Vorgehens und der Regeln zur Transkribierung und Kodierung des Datenmaterials in Abschnitt 3.2 an. Die zur Datenauswertung genutzten probabilistischen Verfahren sind in Kapitel 3.3 ausführlich dargestellt. Neben den Modelleigenschaften werden u. a. Möglichkeiten zum Vergleich konkurrierender Modelle beschrieben, die im Rahmen dieser Arbeit Anwendung finden.

Kapitel 4 widmet sich der Datenauswertung. Einen ersten Einblick in die Datenstruktur bietet Abschnitt 4.1, indem Ergebnisse zur Schreibung ganzer Wörter berichtet werden. Abschnitt 4.2 bildet den Schwerpunkt dieser Arbeit. Hier werden die psychometrischen Eigenschaften der beiden Testinstrumente beschrieben und miteinander verglichen sowie der Erklärungsgehalt alternativer Modelle für die in den Daten enthaltenden Informationen gegeneinander abgewogen. Detailanalysen mit Hintergrundmerkmalen der Kinder aus ZuRecht zeigen in Abschnitt 4.3 mögliche Zusammenhänge mit der Rechtschreibkompetenz auf. Ein weiterer Zugang zu den Daten wird über Profilanalysen unternommen (Abschnitt 4.4), die sich zunächst auf einzelne Fälle und anschließend auf die gesamte Stichprobe beziehen.

In Kapitel 5 findet eine Beantwortung und Diskussion der in Abschnitt 1.2 formulierten Forschungsfragen auf Basis der zuvor durch die Analysen gewonnenen Erkentnisse statt. In diesem psychometrischen Rahmen werden gleichzeitig mögliche Implikationen zur Verbesserung der Tests abgeleitet. Schließlich stellt das Kapitel, zusammen mit einem Ausblick auf noch bestehende Forschungsdesiderate, den Abschluss der Arbeit dar.

# 2 Erfassung von Rechtschreibleistung mit Kompetenzmodellen

In diesem Kapitel werden Studien zur Erhebung von Rechtschreibkompetenz und die beiden in ZuRecht eingesetzten Orthografietests beschrieben. Zunächst erfolgt in Abschnitt 2.1 ein Überblick über den aktuellen Forschungsstand zum Einsatz und zu Ergebnissen von Rechtschreibtests in Schulleistungsstudien. Damit verbunden ist eine Erläuterung der Kompetenzmodelle im Bereich der Orthografie, die in diesen Studien empirisch überprüft wurden. In den Abschnitten 2.2 und 2.3 werden gutschrift-diagnose und der SRT detailliert dargestellt, indem auf die theoretische Konzeption der beiden Tests (Abschnitte 2.2.1 und 2.3.1) sowie empirische Analyseergebnisse (Abschnitte 2.2.2 und 2.3.2) eingegangen wird. Die Darstellung der linguistischen Kompetenzmodelle und der dort angenommenen Teilkompetenzen umfasst ebenfalls Angaben zu verwendeten Wörtern und Indikatoren der Tests, die zur Operationalisierung genutzt werden (in dieser sowie vorherigen Erhebungen). Theoretische Grundlage des SRT sind die Graphematik nach Eisenberg sowie die didaktische Umsetzung nach Hinney. Daher wird der SRT um eine Beschreibung dieser Ansätze angereichert. Den Abschluss des Kapitels bildet Abschnitt 2.4, bei dem die beiden rechtschriftlichen Tests unter theoretischen Gesichtspunkten miteinander verglichen werden.

## 2.1 Studien zu Leistungsstand und Kompetenzmodellüberprüfung

Im Folgenden werden die groß angelegten Rechtschreiberhebungen im Kontext von KESS, der IQB-Pilotstudie und des IQB-Ländervergleichs 2011 dargestellt. Dabei werden die Rahmenbedingungen der Erhebungen, die genutzten Rechtschreibtests sowie die Befunde zum Kompetenzstand der Kinder erläutert. Die aufgeführten Studien werden berücksichtigt, weil sie – analog zu ZuRecht – die folgenden Kriterien erfüllen:

1. Sie beziehen sich auf die Beschreibung des Kompetenzstandes von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe.
2. Rechtschreibung wird differenziert über Teilkompetenzen (bzw. Strategien oder Fehlerebenen und -kategorien) erfasst. Das bedeutet, dass die Schülerschreibungen nicht nur nach Summe von Wortfehlern ausgewertet werden, sondern qualitative Analysen auf Basis theoretischer Modelle erfolgen.
3. Es findet eine Überprüfung der Kompetenzmodelle der eingesetzten Orthografietests nach aktuellen forschungsmethodischen Ansätzen statt.

Im Rahmen von KESS 4 handelt es sich hierbei um die *Hamburger Schreibprobe* und im Hinblick auf die Studien des Instituts zur Qualitätssicherung im Bildungswesen um die *Aachener Förderdiagnostische Rechtschreibfehler-Analyse*. Die Beschreibung des theoretischen Konzepts und der empirischen Kompetenzmodellüberprüfung der beiden Rechtschreibtests erfolgt im Zusammenhang der Studiendarstellungen.

### 2.1.1 KESS 4

#### Erhebungsrahmen

KESS (*Kompetenzen und Einstellungen von Schülerinnen und Schülern*) ist eine in der Hansestadt Hamburg durchgeführte Längsschnittstudie. Es handelt sich um ein kooperatives Forschungsprojekt unter der Leitung von Wilfried Bos, das von der Behörde für Bildung und Sport in Hamburg in Auftrag gegeben wurde. Neben der Erfassung von Kompetenzständen wurden auch vielfältige Hintergrundinformationen zu den Schülerinnen und Schülern über die Befragung von Kontextpersonen (Eltern, Lehrpersonen, Schulleitung) gesammelt. Die Studie hatte drei Erhebungszeitpunkte. In den Jahrgangsstufen 4, 7 und 8 wurden die Schülerinnen und Schüler mit Hilfe eines Fragebogens befragt sowie in unterschiedlichen Domänen getestet. Die Erhebungen fanden in einem Abstand von zwei Jahren statt (Ende der vierten und achten Klasse sowie Anfang der siebten Klasse). Während die Ausgangslagenerhebung im Jahr 2003 durchgeführt wurde, fand die Abschlusserhebung im Jahr 2007 statt. Ziel war es, Aussagen über die Lernstände und -entwicklung der Schülerinnen und Schüler über eine gesamte Schülerkohorte zu formulieren (Bos, Gröhlich, Guill, Scharenberg & Wendt, 2010, S. 10).

Der erste Erhebungszeitpunkt – KESS 4 – soll im Folgenden beschrieben werden. Die Studie wurde flächendeckend in der vierten Klasse durchgeführt. Es beteiligten sich 263 Schulen, 638 Klassen und 14.110 Schülerinnen und Schüler (Bos, Brose et al., 2006, S. 10 f.). Berücksichtigt wurden in unterschiedlichen Fächern die Lernstände, die die Kinder am Ende der Grundschulzeit aufweisen und mit welchen sie auf die weiterführende Schule wechseln. Die Erhebung der Rechtschreibkompetenz fand mit der Hamburger Schreibprobe (HSP) von Peter May sowie der *Dortmunder-Schriftkompetenz-Ermittlung* (DoSE) von Ilona Löffler und Ursula Meyer-Schepers statt. DoSE bezeichnet die Abkürzung der vormaligen Namensgebung von gutschrift-diagnose, die in Abschnitt 2.2

beschrieben wird, weshalb bei diesem Test im Rahmen von KESS ausschließlich auf die Methodik und Ergebnisse eingegangen wird. In KESS erfolgten keine differenziellen Auswertungen nach dem DoSE-Konzept: „Für die Analyse der Rechtschreibstrategien wurde auf die HSP zurückgegriffen, da eine entsprechende Differenzierung für die DoSE nicht vorliegt“ (May, 2006b, S. 111). Daher können ausschließlich globale Maße für DoSE berichtet werden.

### Ergebnisse auf Ganzwortebene

Der in KESS eingesetzte DoSE-Test umfasst ein Lückensatzdiktat mit 45 zu schreibenden Wörtern. Insgesamt haben 6.727 Schülerinnen und Schüler den Test bearbeitet.[1] Durchschnittlich wurden 23,6 Wörter richtig geschrieben, bei einer Standardabweichung (SD) von 9,9. Daneben wurde die Anzahl richtiger Grapheme ausgewertet. Von den 359 Graphemen wurden im Mittel 333,3 richtig erfasst (SD = 19,3) (May, 2006b, S. 112).

Mit der HSP wurden 6.749 Viertklässler getestet. Ihnen wurden insgesamt 42 Testwörter diktiert, wobei es sich zum Teil um Einzelwörter und zum Teil um ganze Sätze handelt. Neben dem Vorlesen der Wörter wurden diese auch über Bilder im Testheft visualisiert. Die Testwerte der Kinder, die die HSP bearbeitet haben, fallen laut May (2006b, S. 112) im Vergleich zur DoSE etwas besser aus, was in der Schreibung der Sätze, welche neben Substantiven ebenfalls acht „Satzstrukturwörter“ (wie z. B. Artikel) beinhalten, begründet liegt. Im Mittel wurden 29,8 Wörter richtig geschrieben (SD = 8,4), womit dieser Test etwa 20 Prozent mehr korrekte Wortschreibungen als DoSE umfasst. Von den 277 zu verschriftenden Graphemen wurden durchschnittlich 256,1 richtig geschrieben (SD = 18,4). Bei der HSP und bei DoSE beträgt die Lösungshäufigkeit der Graphemtreffer damit jeweils 92 Prozent.

Neben der Auswertung der Schülerergebnisse werden Werte zur Reliabilität und zur Validität, die über die Korrelation mit der Rechtschreibnote bestimmt wird, für beide Tests berichtet, welche sich ausschließlich auf die Wort- und Graphemauswahl beziehen. Für die interne Konsistenz werden 0,92 (Wörter) und 0,96 (Grapheme) für die HSP sowie von 0,92 (Wörter) und 0,95 (Grapheme) für DoSE berichtet. Die Zusammenhänge mit der Rechtschreibnote betragen für die HSP 0,73 (Wörter) und 0,69 (Grapheme) sowie für DoSE 0,71 (Wörter) und 0,63 (Grapheme) (May, 2006b, S. 112).

[1] Es gab zwei Testheftversionen: Etwa die Hälfte der Schülerinnen und Schüler der Stichprobe bearbeitete die DoSE und die andere Hälfte die HSP.

### Konzept der HSP

Im Folgenden wird das Konzept der HSP[2] erläutert, da sich die weiterführenden Ergebnisse in KESS auf die Rechtschreibstrategien, also auf eine qualitative Auswertung der Schülerschreibungen nach den angenommenen Teilkompetenzen der HSP, beziehen. Dabei handelt es sich um differenzielle Analysen zum Kompetenzstand der Schülerinnen und Schüler sowie zu dem Kompetenzmodell.

Das Rechtschreibmodell der HSP lehnt sich an das Schriftspracherwerbsmodell der Entwicklungspsychologin Uta Frith an (May, 2006a, S. 19; May, 2002a, S. 61 ff.). Es basiert auf der Annahme, dass Lesen- und Schreibenlernen wechselseitig aufeinander einwirken und dass dieser Prozess in Entwicklungsstufen untergliedert werden kann. Es werden unterschiedliche, nacheinander auftretende Stufen des Schriftspracherwerbs unterschieden. Den Stufen werden Strategien zugeordnet, die jeweils beim Lesen und Schreiben vorherrschen. Frith (1985, S. 307 ff.) postuliert die *logographische*, die *alphabetische* und die *orthografische Strategie*. May erweitert das Modell, wie in Abbildung 2.1 dargestellt ist, um die *morphematische Strategie* und die *wortübergreifende Strategie*. Beim Übergang von einer Stufe in eine nachfolgende wird die neue Strategie mit bereits erworbenen Strategien verknüpft. Auf diese Weise wird der Erwerb der Schriftsprache vorangetrieben. Alte Strategien gehen jedoch nicht verloren, sondern bleiben parallel erhalten, sodass auf diese auch weiter zugegriffen wird (Scheele, 2006b, S. 56). Beim Durchlaufen der Phasen werden das Lesen und Schreiben abwechselnd weiterentwickelt, indem eine bereits höher entwickelte Strategie des Lesens oder des Schreibens das Ausbilden der Strategie innerhalb der jeweils anderen Domäne fördert (Frith, 1986, S. 218).

In der HSP werden die alphabetische, orthografische, morphematische und wortübergreifende Strategie des Entwicklungsmodells ausgewertet und als „grundlegende Zugriffsweisen von Kindern auf die Schrift“ beschrieben (May, 2008a, S. 102). Sie werden von May (2006a, S. 13) über die folgenden Fähigkeiten charakterisiert[3]:

- „Alphabetische Strategie, d. h. die Rekonstruktion der Schreibungen aufgrund der eigenen Artikulation
- Orthographische Strategie, d. h. die Modifikation der einfachen Laut-Buchstaben-Beziehung durch spezifischen [sic] Rechtschreibregeln
- Morphematische Strategie, d. h. die Ableitung und Einpassung der Schreibungen von/in die Bedeutungsstrukturen der Sprache
- Wortübergreifende Strategie, d. h. die Beachtung von Wortarten, Syntax und Textmodalitäten beim Rechtschreiben.“

[2] Neben dem Papiertest HSP existiert der Onlinetest „schreib.on“ (http://www.dideon.de, Zugriff am 11.06.2014). HSP und schreib.on (vormals Deutsche Schreibprobe (DSP)) basieren auf demselben Konzept (http://www.dideon.de/konzept2.html, Zugriff am 11.06.2014).

[3] Eine ausführlichere Beschreibung der Strategien findet sich z. B. in May (2008a, S. 102 f.) oder May (2002a, S. 26 ff.).

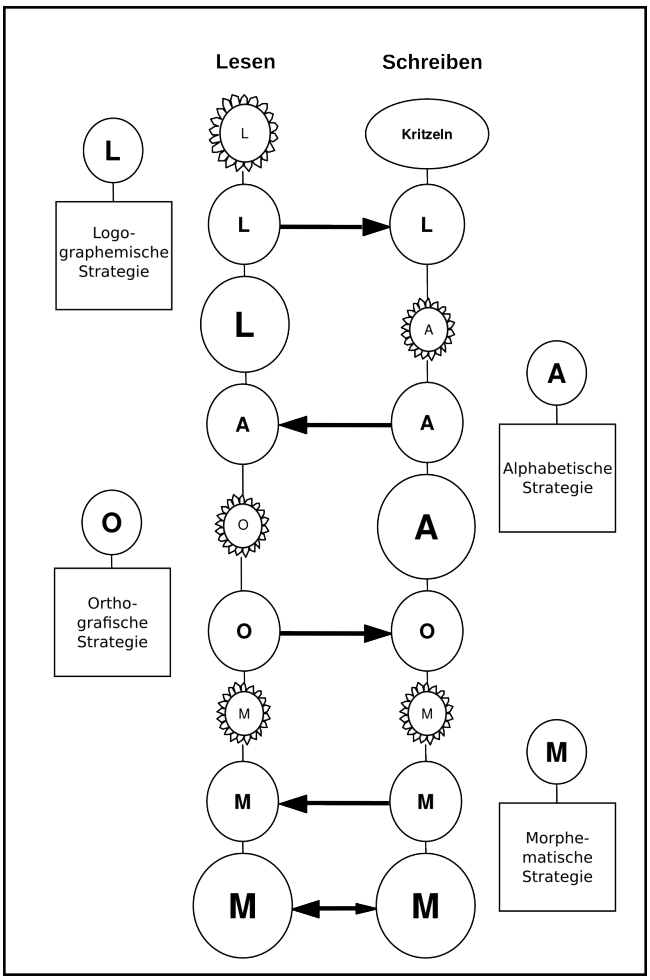



Abbildung 2.1: Das erweiterte Entwicklungsmodell des Schriftspracherwerbs der HSP (nach May, 2008a, S. 105; May, 2002a, S. 62)

Die Differenzierung von orthografischer und morphematischer Strategie stellt eine Besonderheit des HSP-Konzepts dar. Üblich ist die Annahme, dass die Bezugseinheit des Morphems zentraler Bestandteil der orthografischen Stufe ist. Hier werden jedoch morphologische Elemente bzw. Operationen sowohl über orthografische als auch morphematische Zugriffsweisen erklärt, was teilweise eine Unschärfe der beiden Strategiebereiche vermittelt. Beispielsweise wird die Schreibung des Graphems ⟨ß⟩ in Reißverschluss der orthografischen Strategie zugerechnet und die Schreibung des Graphems ⟨d⟩ in Schiedsrichter der morphematischen. In beiden Fällen muss jedoch der Stamm verlängert werden, um zu der richtigen Schreibweise zu gelangen (Herné, 2003, S. 891).

Nach diesen oben genannten in das Entwicklungsmodell eingeordneten Strategien werden die Schreibungen der Kinder ausgewertet und der orthografische Lernstand bestimmt. Dafür werden sogenannte *Lupenstellen* betrachtet. Sie bezeichnen Grapheme innerhalb eines Wortes, die einer Strategie zugeordnet sind und ausgewertet werden. Zusätzlich werden Graphemtreffer, d. h. die Anzahl richtig geschriebener Grapheme, ausgezählt sowie

überflüssige orthografische Elemente und Oberzeichenfehler berücksichtigt (May, 2008a, S. 102). Bei dem Wort Fahrradschloss werden beispielsweise die folgenden Lupenstellen in der HSP4/5 (Version für die Klasse 4) analysiert: Der Anfangsrand ⟨schlo⟩ innerhalb der alphabetischen, das Kürzezeichen ⟨ss⟩ innerhalb der orthografischen sowie die Morphemverbindung ⟨rr⟩ und Auslautverhärtung ⟨d⟩ innerhalb der morphematischen Strategie (May, 2002a, S. 197).

Die Lernentwicklung beim Schriftspracherwerb zeichnet sich nach May (2002a, S. 72 ff.) dadurch aus, dass Kinder zunächst alphabetische und schließlich orthografische sowie morphematische Zugriffsweisen nutzen. Diese Stategien entwickeln sich im weiteren Verlauf der Lernentwicklung zu einer „übergreifenden Gesamtstrategie“. Schriftlernende unterscheiden sich laut May in ihrer Entwicklung zum einen in der Geschwindigkeit der Aneignung der Strategien und zum anderen in der Fähigkeit der Integration der einzelnen Strategien. Mit Hilfe von sogenannten „Strategieprofilen“ werden in der HSP diese Ausprägungen der Strategien ermittelt. Ein ausgeglichenes Profil stellt dabei die gelungende Integration der verschiedenen Strategien dar. „Alphabetische Dominanz“ oder „orthografisch-morphematische Dominanz“ kann auf Störungen im Entwicklungsprozess hindeuten. Hier ist also eine Diskrepanz der Strategiewerte vorhanden. Eine permanente alphabetische Dominanz zeigt, dass diese Strategie unzureichend von orthografischen und morphematischen Zugriffsweisen ergänzt wird. Eine orthografisch-morphematische Dominanz hingegen verdeutlicht, dass die alphabetische Strategie noch nicht hinreichend beherrscht wird, da der Schriftlernende erweiterte Zugriffsweisen, die auf die alphabetische Strategie aufbauen, besser bewältigt. Neben der Gesamtleistung beim Schreiben ist damit die Ausgewogenheit des Strategieprofils zur Diagnostik der Lernentwicklung bei der HSP relevant und damit ein Kennzeichen für stärkere bzw. schwächere Rechtschreiber. (May, 2002a, S. 49 ff., 72 ff.)

Das lernpsychologische Kompetenzmodell der HSP weist die folgenden Hauptmerkmale auf: Es formuliert konkrete Aussagen über die Abfolge von Schritten bei der Rechtschreibentwicklung, indem es dominante Rechtschreibstrategien voneinander unterscheidet. Es handelt sich aber um keine strikt zu trennenden Einzelstrategien, sondern um sich wechselseitig aufeinender bezogende und entwickelnde Zugriffsweisen, die sich zu einer komplexen Gesamtstrategie ausbilden. May differenziert die alphabetische, orthografische, morphematische und wortübergreifende Strategie. Um den rechtschriftlichen Entwicklungsstand eines Kindes zu ermitteln, werden die Schreibprodukte, basierend auf den für eine Strategie relevanten Lupenstellen sowie der Auszählung von Graphemtreffern, ausgewertet und ein Strategieprofil erstellt. Anhand des Profils wird bestimmt, ob eine erfolgreiche Beherrschung und Integration der Einzelstrategien vorliegt.

### Ergebnisse zu den HSP-Strategiebereichen

Für die Darstellung der KESS-Ergebnisse wurde die Stichprobe in drei Schülergruppen differenziert. Die erste Gruppe umfasst 958 Schülerinnen und Schüler, die mindestens eine Standardabweichung über dem Mittelwert aller Hamburger Viertklässler liegen und damit

überdurchschnittliche Leistungen erreichen. Die zweite Gruppe besteht aus 4.707 Kindern mit durchschnittlichen Leistungen, die innerhalb einer Standardabweichung oberhalb bzw. unterhalb des Mittelwertes aller Schülerinnen und Schüler der Stichprobe verortet sind. Die dritte Gruppe besteht aus über 1.084 Kindern. Sie weisen unterdurchschnittliche Leistungen auf, die mindestens eine Standardabweichung unter dem Mittelwert liegen (May, 2006b, S. 114).

Die guten Rechtschreiber (Gruppe 1) verschriften 95,9 Prozent der Wörter und 99,1 Prozent der Grapheme richtig. Die alphabetische Strategie beherrschen sie zu 98,5 Prozent, die orthografische Strategie zu 97,8 Prozent und die morphematische Strategie zu 95,5 Prozent. Kinder mit schwachen Rechtschreibleistungen (Gruppe 3) schreiben 36,3 Prozent der Wörter und 81 Prozent der Grapheme richtig. Damit ist nur gut ein Drittel aller Wörter korrekt geschrieben und etwa jedes fünfte Graphem falsch oder fehlend. Die alphabetische Strategie bereitet den Kindern die wenigsten Probleme. Es werden 70,7 Prozent der Lupenstellen dieses Strategiebereichs korrekt geschrieben. Die orthografische und morphematische Strategie werden dagegen nur zu etwa 40 Prozent beherrscht. Die Schülerinnen und Schüler dieser Gruppe haben „erhebliche Schwierigkeiten, orthografische Regeln und Prinzipien zu verstehen und anzuwenden. Sie können die morphematische Struktur der Wörter erst ansatzweise analysieren." (May, 2006b, S. 117) Die mittleren Rechtschreiber erreichen folgende Werte: 73,8 Prozent richtig geschriebene Wörter, 93,7 Prozent Graphemtreffer, 90,2 Prozent in der alphabetischen, 80,8 Prozent in der orthografischen und 75,5 Prozent in der morphematischen Strategie.

Zusammenfassend lässt sich ein heterogenes Kompetenzniveau beobachten: „Die Lernstände der Kinder gehen am Ende der Grundschulzeit sehr weit auseinander. Während schwache Rechtschreiber lediglich die Lautstrukturen einfach aufgebauter Wörter wiedergeben können, beherrschen fortgeschrittene Lerner zu diesem Zeitpunkt bereits alle wesentlichen orthographischen Phänomene." (May, 2006b, S. 138) Diesen Befund unterstreichen die dokumentierten Effektstärken[4] ($d$) zwischen den Gruppen 1 und 3. Sie pendeln bei der alphabetischen, orthografischen und morphematischen Strategie um $d = 2,5$ und liegen für die richtig geschriebenen Wörter bei $d = 2,96$ sowie für die Graphemtreffer bei $d = 2,72$ (May, 2006b, S. 114).

### Verortung der Ergebnisse aus KESS in weiteren Studien

Bei dem ersten Erhebungszeitpunkt von KESS handelt es sich um eine Querschnittsstudie. Dennoch können die Ergebnisse mit IGLU-E 2001, LAU-5[5] sowie dem Förderprojekt PLUS[6] in Beziehung gesetzt werden. Die Referenzwerte aus IGLU basieren auf den

[4] Nähere Erläuterungen zu Effektstärken beinhaltet Abschnitt 4.3.

[5] LAU-5 bezeichnet die *Lernausgangslagenerhebung* im Schuljahr 1996/1997 in Klassenstufe 5 (Lehmann, Peek & Gänsfuß, 2011). LAU-5 bildet den Auftakt einer in zweijährigem Abstand bis zur 13. Klassenstufe durchgeführten Untersuchung.

[6] PLUS (*Projekt Lesen und Schreiben für alle*) wurde im Zeitraum von 1994 bis 1999 von der ersten bis zur vierten Klasse in Hamburger Schulen durchgeführt. Ziel war die Förderung von Kindern durch

Ergebnissen zur Schreibung ganzer Wörter in DoSE. Ausführliche Ergebnisdarstellungen von DoSE in IGLU 2001 sind in Abschnitt 2.2.2 zusammengestellt.

Bei einem Vergleich der Testergebnisse aus der Hansestadt mit dem in IGLU ermittelten bundesweiten Durchschnitt zeigt sich ein Rückstand der Hamburger Viertklässler. Im Mittel über ganz Deutschland werden 25,6 Wörter (SD = 9,0) und in Hamburg im Durchschnitt 23,6 Wörter (SD = 9,9) richtig geschrieben (Bos, Pietsch & Stubbe, 2006, S. 76). Einen landesweiten Vergleich, der über die Anzahl an Graphemtreffern erfolgt, erlauben die beiden früheren Hamburger Untersuchungen LAU sowie die Voruntersuchung zu PLUS. Die PLUS-Voruntersuchung fand im Jahr 1993/1994 in zufällig ausgewählten vierten Grundschulklassen statt. Die damals getesteten 4.020 Schülerinnen und Schüler schneiden marginal schlechter ab als die KESS-Kinder, was die Effektstärke von $d = 0,08$ veranschaulicht. Für PLUS beträgt der Mittelwert 255,5 und die Standardabweichung 18,4, während in KESS der Mittelwert bei 256,9 und die Standardabweichung bei 17,9 liegen. Die 2.350 Fünftklässler, die im Jahr 1996 an der Lernausgangslagenerhebung teilgenommen haben, erzielen im Vergleich zur Stichprobe in KESS schlechtere Leistungen ($d = 0,21$). Durchschnittlich werden 253,2 Grapheme richtig geschrieben; die Streuung hat einen Wert von 20,2 (May, 2006b, S. 120).

Zusammengefasst lässt sich festhalten, dass KESS Defizite in der Rechtschreibkompetenz der Kinder und heterogene Lernstände innerhalb der vierten Klassen in Hamburg aufdeckt. Zwar gibt es eine leistungsstarke Gruppe von Rechtschreibkönnern, der allerdings eine große Gruppe von Kindern, die lediglich die Lautstrukturen einfach aufgebauter Wörter wiedergeben können, gegenübersteht. In mehreren Strategiebereichen erreichen sie weniger als die Hälfte an richtigen Lösungen. Die Leistungen aller Hamburger Schülerinnen und Schüler haben sich seit den Lernstandserhebungen vor zehn Jahren verbessert. Sie liegen aber dennoch weiterhin unter dem bundesweiten Durchschnitt, wie der Vergleich mit IGLU-E 2001 erbrachte (May, 2006b, S. 138).

### Empirische Überprüfung des HSP-Kompetenzmodells

Die Überprüfung des HSP-Strategiekonzepts erfolgte nicht im Rahmen von KESS, aber mit den in dieser Studie gewonnenen Daten. May nutzte sie, um Interkorrelationen und Zuverlässigkeiten zu bestimmen. Die Werte sind in Tabelle 2.1 abgetragen und beziehen sich auf die alphabetische, orthografische und morphematische Strategie. Die Zusammenhangswerte basieren auf Produkt-Moment-Korrelationen und sind nicht mit latenten Korrelationen zu verwechseln (OECD, 2014, S. 230; OECD, 2012, S. 194). Sie nehmen Werte von 0,69 zwischen der alphabetischen und orthografischen Strategie, von 0,66 zwischen der alphabetischen und morphematischen Strategie sowie von 0,79 zwischen der orthografischen und morphematischen Strategie an. Die Korrelation zwischen der orthografischen und der morphematischen Strategie fällt damit am höchsten aus und könnte die obige Ausführung zu der zum Teil unscharfen theoretischen Trennung widerspiegeln. Die Reliabilität des

Schriftsprachberaterinnen und Schriftsprachberater im Bereich des Lesens und Schreibens (May, 2001a; May, 2000).

| Strategien | (1) | (2) | (3) |
|---|---|---|---|
| (1) Alphabetische Strategie | | | |
| (2) Orthografische Strategie | 0,69 | | |
| (3) Morphematische Strategie | 0,66 | 0,79 | |
| Interne Konsistenz (Cronbachs Alpha) | 0,85 | 0,93 | 0,88 |

Tabelle 2.1: Interkorrelationen und Reliabilitäten der HSP-Strategien (in Anlehnung an May, 2008a, S. 116)

Tests wird über die Angabe von Cronbachs Alpha bestimmt. Die $\alpha$-Koeffizienten liegen für die Strategiebereiche bei 0,85, 0,93 und 0,88. Die Reliabilitäten sind damit als mittelmäßig bis hoch einzuschätzen (Bortz & Döring, 2006, S. 199).

Ebenfalls erfolgte eine mehrdimensionale Raschskalierung[7] über die drei Strategiebereiche. Bei der Itemanalyse waren keine Lupenstellen (Items) auffällig, sodass keine Items eliminiert werden mussten, berichtet May (2008a, S. 117). Die Kriterien für die Analyse werden hier nicht genannt. Die Visualisierung der Gegenüberstellung der Lage und Streuung von Itemschwierigkeiten und Personenfähigkeiten zeigt eine deutliche Verschiebung der Personenfähigkeitsparameter an. Die Items sind im unteren bis mittleren Leistungsbereich verortet und damit eher zu leicht. Neben der dreidimensionalen Skalierung wurde ein eindimensionales Modell[8] berechnet, bei dem keine Strategien voneinander unterschieden, aber alle Lupenstellen auf einer einzigen Dimension abgebildet werden. Die beiden Modelle werden hinsichtlich des Deviance-Werts, der Iterations- sowie Freiheitsgradanzahl verglichen. Konkrete Werte werden von May nicht dokumentiert. Als Ergebnis des Vergleichs hält er fest, dass das dreidimensionale Modell angemessener die Datenstruktur wiedergibt als das eindimensionale Modell (May, 2008a, S. 117 f.).

[7] Das Raschmodell ist in Abschnitt 3.3.2 dargestellt.

[8] Ausführungen zum eindimensionalen Modell bzw. Generalfaktormodell beinhaltet Abschnitt 4.2.1.

### 2.1.2 IQB-Pilotstudie zu den Bildungsstandards

#### Erhebungsrahmen

Die Pilotstudie (oder auch Evaluationsstudie) zu den Bildungsstandards für den Primarbereich wurde vom Institut zur Qualitätssicherung im Bildungswesen (IQB) durchgeführt. Das IQB wurde von der Kultusministerkonferenz beauftragt, die Bildungsstandards zu präzisieren, weiterzuentwickeln, sie mit Tests zu operationalisieren sowie das Erreichen durch Ländervergleiche zu überprüfen. Ziel der Pilotstudie ist die Überprüfung der Tests zur Evaluierung der Bildungsstandards, um geeignete Items auszuwählen, Kompetenzen zu modellieren und zu messen sowie Kompetenzstufenmodelle zu generieren (Granzer et al., 2009, S. 7). Bildungsmonitoring findet im anschließenden Ländervergleich statt (vgl. Abschnitt 2.1.3).

Die Item- und Testentwicklung betraf die Fächer Deutsch und Mathematik. Innerhalb der Domäne Deutsch wurde die Rechtschreibung von 3.480 Schülerinnen und Schülern in Klasse 3 und 4 im Frühjahr 2006 erfasst (Böhme & Bremerich-Vos, 2009, S. 340). Dafür wurden Lückensatzdiktate verwendet, die sich auf vier Aufgabenhefte aufteilen, wobei einige Kinder mehrere Aufgabenhefte bearbeitet haben. Im Rahmen dieser Arbeit werden ausschließlich die Leistungsergebnisse der Viertklässler berichtet.

#### Ergebnisse auf Ganzwortebene

Mit dem ersten Heft wurden 537, mit dem zweiten Heft 518, mit dem dritten Heft 530 und mit dem vierten Heft 549 Viertklässler getestet. Die einzelnen Aufgabenhefte setzten sich jeweils aus zehn Sätzen, in die 20 Testwörter einzutragen waren, zusammen und benötigten jeweils 20 Minuten Bearbeitungszeit (Böhme & Bremerich-Vos, 2009, S. 340). Sie zeigen einen relativ ähnlichen Schwierigkeitsgrad. Es werden folgende Lösungshäufigkeiten angegeben (in aufsteigender Reihenfolge der Aufgabenhefte): 64,7 Prozent, 64,2 Prozent, 66,4 Prozent und 67,0 Prozent Richtigschreibungen auf Wortebene (Böhme & Bremerich-Vos, 2009, S. 342). Die über eine Raschskalierung ermittelten EAP/PV-Reliabilitäten[9] betragen für das erste Heft 0,83, für das zweite und dritte Heft 0,84 sowie für das vierte Heft 0,77 (Böhme & Bremerich-Vos, 2009, S. 347). Ebenso wurden die einzelnen Analysestellen im Wort als ein eindimensionales Modell skaliert. Die Zuverlässigkeit der Messung ist für die Items der ersten beiden Aufgabenhefte mit den zuvor genannten Ergebnissen auf Ganzwortebene identisch. Beim dritten Heft unterscheidet sie sich nur leicht mit 0,83, während die Reliabilität für das vierte Heft auf 0,72 absinkt. Die Lupenstellen wurden auf Basis der Aachener Förderdiagnostischen Rechtschreibfehler-Analyse (AFRA) bestimmt, die im Folgenden beschrieben wird.

[9] Ausführungen zur EAP/PV-Reliabilität umfasst Abschnitt 4.2.3.

| | | |
|---|---|---|
| **Phonem-Graphem-Korrespondenz** | | |
| **BF** | **Buchstaben-Form**<br>spiegelbildlich, unvollständig oder unleserlich geschriebener Buchstabe | *Hanb (Hand)<br>*überQueren (überqueren)<br>*Manner (Männer) |
| **GA** | **Graphem-Auswahl**<br>Auswahl eines Graphems, das keine lauttreue Verschriftung des betreffenden Phonems darstellt | *Prei (Brei)<br>*schlümm (schlimm) |
| **GF** | **Graphem-Folge**<br>Auslassung oder Hinzufügung eines Graphems oder Vertauschung der Reihenfolge von Graphemen | *Wurt (Wurst)<br>*Fabirk (Fabrik) |
| **SG+** | **Spezielle Grapheme (Mehrheit)**<br>Fehler bei <ch>, <f>, <k>, <ng>, <r>, <s> oder <sch> | *vangen (fangen)<br>*Schrancke (Schranke)<br>*Roße (Rose) |
| **SG-** | **Spezielle Grapheme (Minderheit)**<br>Fehler bei <v> oder <ß> | *foll (voll)<br>*giesen (gießen) |
| **SV+** | **Spezielle Verbindungen (Mehrheit)**<br>Fehler bei <au>, <ei>, <nk>, <sp>, <st>, <x> oder <z> | *Ongkel (Onkel)<br>*schpielen (spielen) |
| **SV-** | **Spezielle Verbindungen (Minderheit)**<br>Fehler bei <ai>, <chs>, <pf> oder <qu> | *Keiser (Kaiser)<br>*erwaxen (erwachsen)<br>*Strumf (Strumpf) |
| **FW** | **Fremdwort-Grapheme**<br>Nichtbeachtung einer fremdsprachlichen Phonem-Graphem-Korrespondenz | *Teater (Theater)<br>*Compjuter (Computer) |
| **Vokalquantität** | | |
| **LI+** | **Langes i (Mehrheit)**<br>Fehler bei der Schreibung von /i:/ als <ie> | *siben (sieben)<br>*Bihne (Biene)<br>*spillen (spielen) |
| **LI-** | **Langes i (Minderheit)**<br>Fehler bei der Schreibung von /i:/ als <i>, <ih> oder <ieh> | *Tieger (Tiger)<br>*siet (sieht)<br>*Apfelsinne (Apfelsine) |
| **LV+** | **Lange Vokale (Mehrheit)**<br>Fehler bei der Schreibung eines ungekennzeichneten Langvokals (außer /i:/) | *lohben (loben)<br>*gebben (geben) |
| **LV-** | **Lange Vokale (Minderheit)**<br>Fehler bei der Schreibung eines durch Dehnungs-h oder Doppelvokal gekennzeichneten langen Vokals | *faren (fahren)<br>*Mohs (Moos)<br>*Stull (Stuhl) |
| **$KV_{Ø}$+** | **Kurzvokale ohne Kennzeichnung (Mehrheit)**<br>Fehler bei der Schreibung eines ungekennzeichneten kurzen oder unbetonten Vokals | *Kappelle (Kapelle)<br>*Hefft (Heft)<br>*Stuhnde (Stunde) |
| **$KV_{D}$+** | **Kurzvokale mit Kennzeichnung (Mehrheit)**<br>Fehler bei der Schreibung eines durch Doppelkonsonant bzw. <ck> oder <tz> gekennzeichneten Kurzvokals | *Bal (Ball)<br>*komt (kommt) |
| **KV-** | **Kurzvokale (Minderheit)**<br>Nichtbeachtung einer irregulären Schreibung eines kurz gesprochenen oder unbetonten Vokals | *ann (an)<br>*Baterie (Batterie) |
| **Morphologie** | | |
| **MS** | **Morphologische Segmentierung**<br>fehlerhafte Verschriftung eines Morphemanschlusses | *Fahrad (Fahrrad)<br>*träumpt (träumt)<br>*kent (kennt) |
| **MD** | **Morphem-Differenzierung**<br>korrekte Schreibung eines gleich oder ähnlich lautenden Morphems | *Warheit (Wahrheit)<br>*Erdbären (Erdbeeren) |
| **UM** | **Unselbstständige Morpheme**<br>fehlerhafte Verschriftung eines unselbstständigen Morphems | *Ferbot (Verbot)<br>*lustisch (lustig) |
| **KA+** | **Konsonantische Ableitung (Mehrheit)**<br>Nichtbeachtung der Verlängerungsregeln bei <b/p>, <d/t>, <g/k>, <h> oder <s/ß> | *runt (rund)<br>*Fleis (Fleiß)<br>*hept (hebt) |
| **KA-** | **Konsonantische Ableitung (Minderheit)**<br>Nichtbeachtung der irregulären Schreibung eines Konsonanten im Endrand eines Morphems | *Jugent (Jugend)<br>*Opst (Obst) |
| **VA+** | **Vokalische Ableitung (Mehrheit)**<br>Nichtbeachtung der Ableitungsregeln bei <e/ä> bzw. <eu/äu> | *Menner (Männer)<br>*gärn (gern)<br>*Bäute (Beute) |
| **VA-** | **Vokalische Ableitung (Minderheit)**<br>Nichtbeachtung einer irregulären <e>- oder <ä>- Schreibung | *Ältern (Eltern)<br>*Seule (Säule) |
| **Syntax** | | |
| **GK+** | **Groß- und Kleinschreibung (Mehrheit)**<br>fälschliche Großschreibung eines Wortes | ein *Großer Ball<br>Das *Mag ich nicht |
| **GK-** | **Groß- und Kleinschreibung (Minderheit)**<br>fälschliche Kleinschreibung eines Wortes | ein kleiner *hund<br>er kommt aus *aachen<br>das *gefühl |
| **ZG** | **Zusammen- und Getrenntschreibung**<br>fälschliche Getrennt- oder Zusammenschreibung | beim *Obst Schälen<br>*Olympischespiele |

Tabelle 2.2: Die Fehlersystematik der AFRA (nach Herné & Naumann, 2009, Anhang Fehlerkategorien (Kurzbeschreibung))

### Konzept der AFRA

Bei der AFRA handelt es sich um eine förderdiagnostische Fehleranalyse von Karl-Ludwig Herné und Carl Ludwig Naumann (Herné & Naumann, 2009, S. 5; Herné, 1993, S. 323). Schreibungen werden qualitativ ausgewertet, indem – in Anlehnung an die Begriffsverwendung von May – Lupenstellen betrachtet werden. Ergebnis der qualitativen Auswertung ist die Anfertigung individueller Fehlerprofile von Schülerinnen und Schülern. Die AFRA kann, durch die Verfügbarkeit von Auswertungsrastern und -bögen, auch auf andere Rechtschreibtests (wie z. B. die HSP) angewendet werden (Herné & Naumann, 2009, S. 4 f.). Herné und Naumann (2009, S. 42 ff.) bieten dazu einen breiten Fundus an Erläuterungen, Arbeitshilfen und praktischen Beispielen zur Auswertung nach den AFRA-Kategorien an.

Das Konzept ist charakterisiert durch eine linguistische Kategorisierung von Rechtschreibfehlern in 16 Hauptkategorien auf vier Fehlerebenen: *phonologische Ebene*, *lange und kurze Vokale*, *morphologische Ebene* sowie *syntaktische Ebene*. Die 16 Fehlerkategorien teilen sich weiter in 25 Unterkategorien auf, da sie in *Mehrheits- und Minderheitsschreibungen* differenziert werden (Herné & Naumann, 2009, S. 7). Die Definition von Mehrheits- und Minderheitsschreibungen[10] bildet ein weiteres wichtiges Strukturierungsprinzip in der AFRA. Es basiert auf sprachstatistischen Besonderheiten, die die Unterscheidung in Regel- und Ausnahmefälle widerspiegeln (Herné, 2003, S. 893; Herné & Naumann, 2009, S. 8). Für jedes Phonem werden Realisierungswahrscheinlichkeiten der Allographe über statistische Mehrheitsverhältnisse ausgegeben (Herné & Naumann, 2009, S. 8; Herné, 1993, S. 321). Beispielsweise wird das /f/-Phonem im Geschriebenen in 75 Prozent der Fälle als ⟨f⟩ wiedergegeben, in 13 Prozent als ⟨ff⟩, in 11 Prozent als ⟨v⟩ sowie in 1 Prozent der Fälle als ⟨ph⟩. Daher stellt ⟨f⟩ die Mehrheitsschreibung dar (Herné, 1993, S. 321 f.). Die Häufigkeitsverteilungen von Langvokalschreibungen werden ebenfalls in Mehrheits- und Minderheitsschreibungen eingeteilt. Demnach sind Schreibungen mit langem Vokal in 87,8 Prozent der Fälle unmarkiert, erfolgen in 11,6 Prozent der Fälle über die Setzung eines Dehnungs-h sowie in 0,6 Prozent der Fälle über Vokalgraphemverdopplung. Der Anfangsunterricht sollte nach Herné (1993, S. 322) daher auf den mehrheitlichen Regularitäten aufbauen. Mehrheitsschreibungen werden in AFRA durch ein „+“ hinter dem Kategorienkürzel gekennzeichnet, Minderheitsschreibungen durch ein „-“. AFRA bietet damit Hinweise auf unterschiedliche Gewichtungen von Rechtschreibfehlern bzw. relativiert diese (Herné & Naumann, 2009, S. 17 f.; Herné, 1993, S. 321 ff.). Die Kenntnis der Mehrheitsverhältnisse ist für den Diagnostiker als auch den Schriftlernenden hilfreich (Herné, 2003, S. 893). Eine Beschreibung der einzelnen Minderheitenschreibungen erfolgt im Rahmen eines Exkurses in Abschnitt 2.4.3.

Aus den genannten Ebenen und den 16 Hauptfehlerkategorien, die auf Basis der Mehrheits- und Minderheitsschreibungen in insgesamt 25 Fehlerkategorien untergliedert werden, ergibt sich die in Tabelle 2.2 dargestellte Fehlersystematik. Die Kategorien sind auf Basis der Ebenen geordnet und werden durch Beispiele ergänzt. Bei den Kategorien handelt es sich

[10] Bei G. Thomé und D. Thomé (2010, S. 9 f.) findet sich eine ähnliche Unterscheidung zwischen *Basisgraphemen* und *Orthographemen*.

um Kurzbeschreibungen, die jeweils unterschiedliche Verstöße gegen rechtschreibliche Prinzipien und Regeln zusammenfassen (Herné, 1993, S. 323). In der Fehlersystematik fällt u. a. die Stellung der Vokalquantität auf, über die eine eigene Ebene mit sieben Fehlerkategorien gebildet worden ist. Durch die hohe „Fehlerverlockung“ bei der Schreibung von langen und kurzen Vokalen wurde diesem Bereich eine Sonderstellung eingeräumt (Herné, 1993, S. 323). Bei der Kategorie MD geht es um die semantische Unterscheidung von Homophonen (Morpheme mit gleicher phonologischer, aber unterschiedlicher graphematischer Form, wie z. B. bei mehr und Meer) und bei UM um die Schreibung von Derivations- und Flexionsmorphemen sowie Fugenelementen (Herné & Naumann, 2009, S. 14). Bei der syntaktischen Ebene wird die Kleinschreibung als Normalfall angesehen, weshalb großzuschreibende Wörter in die Kategorie GK- fallen, wie z. B. Großschreibung des Satzanfanges oder von Konkreta und Abstrakta (Herné & Naumann, 2009, S. 16). In die Kategorie GK- werden u. a. alle nicht lexikalisierten Substantivierungen einsortiert.

Bei Fahrradschloss, einem Testwort der HSP 4/5 (vgl. Abschnitt 2.1.1), erfolgt eine Auswertung nach zwölf AFRA-Kategorien. Insgesamt werden in Herné und Naumann (2009, Anhang Auswertungsschemata) die folgenden Lupenstellen analysiert[11]:

| | | |
|---|---|---|
| BF | → | * |
| GA | → | * |
| GF | → | * |
| SG+ | → | ⟨f⟩, ⟨r[r]⟩, ⟨sch⟩ |
| LV+ | → | ⟨ad⟩ |
| LV- | → | ⟨ahr⟩ |
| $KV_D$+ | → | ⟨oss⟩ |
| MS | → | ⟨r'r⟩, ⟨d'sch⟩ |
| MD | → | ⟨*fach⟩, ⟨*rat⟩ |
| KA+ | → | ⟨d⟩ |
| GK- | → | ⟨F⟩ |
| ZG | → | ⟨rr⟩, ⟨ds⟩ |

Analog zur HSP (s. o.) wird die Verschriftung von ⟨rr⟩ und ⟨d⟩ unter einen morphologischen Bereich gefasst. In der HSP erfolgt die Analyse von ⟨ss⟩ allerdings unter der orthografischen Strategie und in der AFRA von ⟨oss⟩ unter der Kategorie Vokalquantität. Zudem fällt die Anzahl ausgewiesener Lupenstellen in der AFRA höher aus. So findet hier – im Gegensatz zur HSP – z. B. eine Betrachtung der Mehrheitsschreibung des Graphems ⟨f⟩ (SG+), des Dehnungs-h (Kategorie LV-), des ungekennzeichneten langen Vokals ⟨a⟩ (LV+), der Groß- und Klein- (GK-) sowie der Zusammen- und Getrenntschreibung (ZG) statt. Allerdings ist dabei zu berücksichten, dass in der HSP neben den Lupenstellen die Graphemtreffer erfasst und ausgezählt werden.

[11] Morphemgrenzen werden durch ein Apostroph dargestellt; bei BF, GA und GF befindet sich bei den Items im Auswertungsschema immer ein Stern und mehrfache Fehler innerhalb einer der Kategorien werden nur einmal gekennzeichnet (Herné & Naumann, 2009, S. 47).

Die Ebenen und Kategorien basieren auf der Annahme der Autoren, dass die Orthografie systematisch ist (Herné & Naumann, 2009, S. 6, 16). Visualisiert haben sie die Konzeption in dem „Haus der Orthografie“: „Das ‚orthografische Haus‘ formuliert die Tatsache anschaulich, dass die Orthografie des Deutschen ein ziemlich geordnetes Ganzes ist.“ (Herné & Naumann, 2009, S. 16) Auf vier Etagen werden Bereiche der Orthografie abgebildet, wobei die Etagen gleichzeitig die Anforderungen an den Schriftlernenden darstellen (Herné & Naumann, 2009, S. 7). Das Fundament bildet das Erdgeschoss mit der sprachlichen Ebene der Laute, das Zwischengeschoss ist für die Vokaldauer reserviert, die Wortbausteine belegen das 1. Obergeschoss und im 2. Obergeschoss findet sich die Ebene des Satzes. Der Dachstuhl beinhaltet eine „Rumpelkammer“ mit Ausnahmen und Raritäten (Herné & Naumann, 2009, S. 16).[12]

Für die IQB-Pilotstudie wurden die Fehlerkategorien reduziert bzw. zusammengefasst. Die Mehrheits- und Minderheitsschreibungen bei der Kategorie SG wurden aufgelöst, sodass die erste Fehlerkategorie *spezielle Grapheme und Graphemverbindungen* (SG) lautet. Die weiteren differenzierten Fehlerkategorien aus der Ebene der Phonem-Graphem-Korrespondenz werden nicht übernommen. Im Bereich der Vokalquantität werden VL+ und VL- adaptiert, wobei die Kategorien LI+ und LI- entsprechend subsumiert werden. Die zweite und dritte Kategorie lauten demnach *Vokallänge in der Mehrheit der Fälle* (VL+) und *Vokallänge in der Minderheit der Fälle* (VL-). Die Kategorie $KV_D+$ wird bei der Pilotstudie übernommen und hier als vierte Kategorie unter der Bezeichnung *Vokalkürze* (VK) geführt, sodass keine gesonderte Betrachtung der Kategorien $KV_Ø+$ und KV- erfolgt. Auf der Ebene der Morphologie wird UM in die fünfte Kategorie *häufige Morpheme* (HM) transferiert, sowie MD in die sechste Kategorie *Morphemgrenze* (MG), worunter auch das Fugenelement ⟨s⟩ fällt. Die Mehrheits- und Minderheitsschreibungen bei KA und VA werden zusammengefasst. Somit stellen *vokalische Ableitung* (VA) und *konsonantische Ableitung* (KA) die siebte und achte Kategorie dar. Hier ist zu berücksichtigen, dass Fehlerstellen, bei denen durch Verlängerung des Wortes die Schreibung deutlich wird (z. B. lieblich mit ⟨ch⟩ oder ⟨g⟩), in die Kategorie KA und nicht in HM einsortiert werden. Die Ebene Syntax wird durch die neunte Kategorie *Groß- und Kleinschreibung* (GK) abgebildet. Auch hier erfolgt keine Unterscheidung zwischen Mehrheits- und Minderheitsschreibungen. Eine letzte Kategorie wurde in der Pilotstudie unter dem Namen *anderer Fehler* (AF) eingeführt, über die Fehler erfasst werden sollen, die die zuvor genannten neun Kategorien nicht beinhalten (Böhme & Bremerich-Vos, 2009, S. 338 f.). Im Folgenden sind die neun Kategorien der IQB-Pilotstudie (AF wird bei der Auswertung nicht berücksichtigt), die in Orientierung an die AFRA-Fehlerkategorien gebildet wurden, maßgeblich.

Die Auswahl der Testwörter in der Evaluationsstudie erfolgte mithilfe des Orientierungswortschatzes von Naumann (1999) sowie nach den folgenden Kriterien: Es sollten hauptsächlich Wörter aus den Bereichen spezielle Grapheme (SG) und spezielle Verbindungen (SV) vertreten sein, da die einfachen Phonem-Graphem-Korrespondenzen von der Mehrheit der Schülerinnen und Schüler im zweiten Schuljahr beherrscht werden. Dagegen stellen Vokallänge und -kürze bis in die weiterführende Schule einen Problembereich dar,

[12] Das Haus ist u. a. in Herné und Naumann (2009, S. 17) abgebildet.

weshalb Wörter insbesondere aus den Bereichen VL+ (u. a. Schreibung von ⟨ie⟩) und VL- (z. B. Dehnungs-h) sowie Sonderschreibweisen für Doppelkonsonanten (⟨tz⟩, ⟨ck⟩) im Test vorhanden sein sollten. Vorkommen sollten ebenfalls Komposita, Derivationen (häufige Affixe) und Flexionsformen (KA+ und VA+), sowie im Bereich der Großschreibung (G/K) Konkreta, Substantive mit Artikel und typischem Suffix, Abstrakta und Substantivierungen (Böhme & Bremerich-Vos, 2009, S. 339).

### Empirische Überprüfung der AFRA-Fehlerkategorien

In der Pilotstudie zu den Bildungsstandards wird für den Bereich der Rechtschreibung u. a. das Untersuchungsziel formuliert, die theoretisch abgeleiteten didaktischen Annahmen zur Differenzierung der Fehlerkategorien über empirische Befunde zu stützen (Böhme & Bremerich-Vos, 2009, S. 331). Die in diesem Zusammenhang berichteten Analyseergebnisse beziehen sich alle auf das zweite Aufgabenheft, welches insgesamt 1.050 Schülerinnen und Schüler der dritten und vierten Klassenstufe bearbeitet haben (Böhme & Bremerich-Vos, 2009, S. 343). Das berechnete Modell weist entsprechend der Fehlerkategorien die neun Dimensionen Spezielle Grapheme (SG), Vokallänge in der Mehrheitsschreibung (VL+), Vokallänge in der Minderheitenschreibung (VL-), Vokalkürze (VK), Häufige Morpheme (HM), Morphemgrenzen (MG), Vokalische Ableitungen (VA), Konsonantische Ableitungen (KA) sowie Groß- und Kleinschreibung (GK) auf. Die Autoren berichten von Konvergenzproblemen aufgrund der im Verhältnis zur Itemanzahl vielen Dimensionen. Auch bei einem sechsdimensionalen Modell, bei dem die Dimensionen VL+ und VL-, HM und MG (Morphologie) sowie VA und KA (Ableitungen) zusammengelegt wurden, blieben diese erhalten (Böhme & Bremerich-Vos, 2009, S. 348). Als Ergebnis der Skalierungen wird von schlechten Modellanpassungen berichtet. Die latenten Interkorrelationen und Reliabilitäten für das Modell mit neun unterschiedenen Dimensionen sind in Tabelle 2.3 dargestellt. Die Maße der EAP/PV-Reliabilität liegen zwischen 0,67 und 0,89. Die Korrelationen auf latenter Ebene pendeln zwischen 0,66 und 0,96 und liegen 13-mal über 0,9.

In einem nächsten Schritt erfolgte eine weitere Reduktion auf drei Dimensionen. Dort stellt SG die erste Fehlerkategorie dar. VL+, VL- und VK, die unter dem Bereich Dehnung und Schärfung (DS) zusammengeführt werden, bilden die zweite Kategorie. Mit HM, MG, VA, KA und GK wurde eine dritte Fehlerkategorie, die mit Morphologie (MO) betitelt wird, eingeführt. Die Reliabilitätsschätzungen dieses Modells liegen bei 0,77 (SG), 0,84 (DS) und 0,80 (MO). Die Korrelationen mit SG betragen für DS und MO jeweils 0,88. Der Zusammenhangswert zwischen MO und DS liegt bei 0,91. Für die Autoren ist dieses Ergebnis, so wie sie schreiben, nicht zufriedenstellend.

Daran anschließend erfolgte eine *Hauptkomponentenanalyse* (PCA, Principal Components Analysis) mit den Daten auf Fehlerebene (eindimensionales Modell mit Lupenstellen). Dabei handelt es sich um ein Verfahren der Faktorenanalyse und damit um eine Methode der Datenreduktion (Borg & Staufenbiel, 2007, S. 201 ff.). Im Mittel konnten rund 30 Prozent der Varianz bei einem extrahierten Hauptfaktor aufgeklärt werden. Der Quotient

| Fehler-kategorien | SG | VL+ | VL- | VK | HM | MG | VA | KA | GK |
|---|---|---|---|---|---|---|---|---|---|
| SG | 0,79 | | | | | | | | |
| VL+ | 0,84 | 0,89 | | | | | | | |
| VL- | 0,79 | 0,97 | 0,88 | | | | | | |
| VK | 0,82 | 0,96 | 0,94 | 0,88 | | | | | |
| HM | 0,91 | 0,92 | 0,91 | 0,90 | 0,87 | | | | |
| MG | 0,78 | 0,95 | 0,91 | 0,93 | 0,87 | 0,78 | | | |
| VA | 0,84 | 0,93 | 0,93 | 0,93 | 0,93 | 0,88 | 0,84 | | |
| KA | 0,66 | 0,74 | 0,74 | 0,68 | 0,75 | 0,67 | 0,68 | 0,67 | |
| GK | 0,72 | 0,73 | 0,72 | 0,76 | 0,69 | 0,68 | 0,70 | 0,68 | 0,67 |

Tabelle 2.3: Latente Interkorrelationen (untere Dreiecksmatrix) und Reliabilitäten (Hauptdiagonale) der AFRA (nach Böhme & Bremerich-Vos, 2009, S. 349)

aus erstem und zweitem Eigenwert beträgt durchschnittlich 6,89, was von den Autoren als deutlicher Hinweis auf Eindimensionalität gewertet wird (Böhme & Bremerich-Vos, 2009, S. 350). Als weiteres Kriterium zur Bestimmung der Faktorzahl wurde die Parallelanalyse gewählt. Dabei werden empirische und simulierte (zufallsbedingte) Eigenwerte miteinander verglichen und diejenigen Faktoren extrahiert, deren Eigenwerte deutlich über den Zufallseigenwerten liegen (Wolff & Bacher, 2010, S. 343 f.). Da der zweite empirische Eigenwert in der IQB-Pilotstudie jeweils den zweiten simulierten Eigenwert übersteigt, spricht dies hingegen eher gegen Eindimensionalität.

Die Autoren halten fest, dass aufgrund der Befunde der verschiedenen Analysen „niedrigdimensionale Lösungen plausibler sind als höherdimensionale“, weshalb mit einem „eindimensionalen Modell zur Beschreibung der Rechtschreibkompetenz gearbeitet werden kann“ (Böhme & Bremerich-Vos, 2009, S. 350). Die Ergebnisse zur Dokumentation der Bildungsstandards beziehen sich demzufolge auf ein Globalmaß.

### 2.1.3 IQB-Ländervergleich 2011

#### Erhebungsrahmen

Die Ländervergleiche werden vom Institut zur Qualitätsentwicklung im Bildungswesen (IQB) durchgeführt. Sie sind an die Stelle der nationalen Ergänzungsstudien von PISA (*Programme for International Student Assessment*) und IGLU getreten und verfolgen das Ziel, das Erreichen der nationalen Bildungsstandards zu überprüfen sowie empirisch fundierte Kompetenzstufenmodelle zu entwickeln (Rabe, 2012, S. 9). Die Kompetenzstufenmodelle wurden auf Basis der Ergebnisse der Pilotstudie (s. o.) sowie auf Grundlage

von fachdidaktischen Kompetenzstrukturmodellen definiert. Dabei werden kontinuierliche Kompetenzskalen in Abschnitte (Kompetenzstufen oder -niveaus[13]) eingeteilt. Die Grenzen der Kompetenzstufen werden von Experten aus der Fachdidaktik, Psychometrie und Bildungsadministration bestimmt. Kompetenzstufenmodelle erlauben eine Zuordnung von Schülerleistungen und Testitems sowie eine Festlegung von *Mindest-, Regel-* und *Maximalstandards* zu den definierten Kompetenzniveaus (Böhme, Richter, Stanat, Pant & Köller, 2012, S. 17). Mindeststandards werden definiert als ein Minimum an vorhandenen Kompetenzen bei den Schülerinnen und Schülern. Regelstandards bzw. Regelstandards plus beziehen sich auf einen Kompetenzstand, der im Einklang mit den KMK-Publikationen steht bzw. darüber liegt. Leistungserwartungen, die die beschriebenen Anforderungen der KMK-Bildungsstandards weit übertreffen, werden als Maximalstandards bezeichnet (Pant et al., 2012, S. 54 f.).

Der Ländervergleich in der vierten Jahrgangsstufe erfolgte im Jahr 2011. Erhoben wurden die Leistungen in Deutsch und Mathematik von 27.081 Schülerinnen und Schülern aus allen Bundesländern (Richter et al., 2012, S. 85, 95). Dabei kam ein Multi-Matrix-Design zum Einsatz, bei dem jedes Kind nur eine Teilmenge an Items beantwortet (Weirich, Haag & Roppelt, 2012, S. 277). Im Fach Deutsch wurden die Kompetenzen in den Bereichen Lesen, Zuhören und Rechtschreibung erfasst (Richter et al., 2012, S. 86). Der länderübergreifende Vergleich bezieht sich auf diese Kompetenzbereiche des Faches Deutsch und eine globale Mathematikleistung. Orthografie wurde in einer nur für Gesamtdeutschland repräsentativen Stichprobe erfasst, die keinen Ländervergleich zulässt (Stanat et al., 2012, S. 171). Ziel bildete damit ausschließlich die Überprüfung des Kompetenzstufenmodells (Richter et al., 2012, S. 86).

Die Wortauswahl zur Überprüfung der Bildungsstandards basiert auf Grundwortschätzen, dem Orientierungswortschatz von Naumann sowie auf Wörtern aus bewährten standardisierten Tests. Ferner wurden hauptsächlich Wörter mit orthografischen Besonderheiten ausgesucht, da Kinder in der vierten Jahrgangsstufe überwiegend Phonem-Graphem-Zuordnungen beherrschen (Böhme & Bremerich-Vos, 2012, S. 30 f.). Eingebettet wurden sie in Lückensatzdiktaten und Korrekturaufgaben. Bei den Korrekturaufgaben sollten vorgegebene Schreibungen korrigiert werden. Hierbei werden passive Aspekte orthografischer Kompetenz getestet, wie die Fehlersensibilität sowie die Kontrolle und Korrektur auf orthografische Richtigkeit. Beide Testformen beinhalten jeweils etwas mehr als 40 Prozent der Orthografie-Items. Bei den verbleibenden 15 Prozent der Testitems handelt es sich um unterschiedliche Aufgabentypen, die insbesondere die Anwendung von Strategiewissen, also von orthografischen und morphematischen Regeln, erfassen (z. B. bei Ableitungsoperationen) (Böhme & Bremerich-Vos, 2012, S. 32 f.). Insgesamt wurden 152 Items eingesetzt (Weirich et al., 2012, S. 278).

[13] Weiterführende Informationen finden sich in Pant, Böhme und Köller (2012, S. 49 ff.).

### Untersuchte Fehlerkategorien und Stufenzuordnungen

Die Rechtschreibleistungen wurden in Orientierung an den Fehlerkategorien der AFRA ausgewertet. Das Kategorienmodell der Evaluationsstudie wurde dabei größtenteils übernommen. Es erfolgte allerdings eine Ergänzung der zwei Kategorien Graphem-Auswahl (GA) und Graphemfolge (GF), die aus der ursprünglichen AFRA-Fehlersystematik stammen (vgl. Tabelle 2.2). Dementsprechend werden die folgenden elf Fehlerkategorien im Ländervergleich voneinander unterschieden: 1. Graphem-Auswahl (GA), 2. Graphemfolge (GF), 3. Spezielle Grapheme (SG), 4. Vokallänge in der Mehrheit der Fälle (VL+), 5. Vokallänge in der Minderheit der Fälle (VL-), 6. Vokalkürze (VK), 7. vokalische Ableitung (VA), 8. konsonantische Ableitung (KA), 9. häufige Morpheme (HM), 10. Morphemgrenze (MG) sowie 11. Groß- und Kleinschreibung (GK) (Böhme & Bremerich-Vos, 2012, S. 31 f.).

Den Fehlerkategorien werden Stufen zugeordnet, da von einem Modell mit stufenweisem Aufbau orthografischer Kompetenzen ausgegangen wird. Die Stufen sind aber keine klar voneinander abgrenzbaren aufeinanderfolgenden Entwicklungsstufen, sondern sukzessiv entfaltende und parallel verlaufende Zugriffsweisen oder Strategien (Böhme & Bremerich-Vos, 2012, S. 29 f.). Die erste Stufe ist die alphabetische, die durch die Beherrschung der Phonem-Graphem-Korrespondenzen auf der Ebene der Basisgrapheme gekennzeichnet ist. Eine zweite Stufe wird als orthografische Stufe bezeichnet. Hier kommen die Kinder mit den Rechtschreibphänomenen Dehnung, Schärfung, Umlautableitung, Auslautverhärtung, „Merkelemente“ (z. B. ⟨v⟩ bei Vater), Stammprinzip, Großschreibung von Substantiven sowie Abgrenzung von Sätzen in Berührung. Unter die alphabetische Stufe werden Fehler bei den Kategorien GA und GF gerechnet. Alle weiteren Kategorien finden sich innerhalb der orthografischen Stufe wieder (Böhme & Bremerich-Vos, 2012, S. 29 ff.).

### Kompetenzstufenmodell Orthografie und Verteilung auf den Stufen

Die Ausgestaltung der Kompetenzstufen lehnt sich an die Lupenstellen der Fehlerkategorien und die angenommenen Stufen bzw. Strategien an. Es werden fünf Kompetenzstufen unterschieden, wobei die erste und letzte Stufe nach unten bzw. oben offen ist:

- Kompetenzstufe I: unter 390 Punkte
- Kompetenzstufe II: 390 bis 464 Punkten (Mindeststandard)
- Kompetenzstufe III: 465 bis 539 Punkte (Regelstandard)
- Kompetenzstufe IV: 540 bis 614 Punkte (Regelstandard plus)
- Kompetenzstufe V: 615 Punkte und mehr (Maximalstandard)

Die Stufenbreite beträgt jeweils 75 Punkte. Den Kompetenzstufen II bis V werden die verschiedenen erzielten Ausprägungen der Bildungsstandards zugeordnet (Bremerich-Vos, Böhme, Krelle, Weirich & Köller, 2012, S. 67). Die Charakterisierung der Kompetenzstufen wird im Folgenden zusammenfassend erläutert, wobei auf das Erreichen der Kompetenzen im Lückensatzdiktat und bei den Korrekturaufgaben eingegangen wird:

Kompetenzstufe I Die Lautstruktur ist durch die Wortschreibungen erkennbar, wobei noch Unsicherheiten bei den Phonem-Graphem-Zuordnungen bestehen. Damit wird die alphabetische Stufe noch nicht sicher beherrscht. Der Anteil der Fehler in den Kategorien dieser Stufe beträgt bei GA 10 Prozent und bei GF etwas über 10 Prozent. Richtig geschrieben werden zudem teilweise Dehnungs-h (VL-), konsonantische Ableitungen in strukturell einfachen Wörtern (KA), Markierungen kurzer Vokale (VK), Morphemgrenzen (MG) sowie Konkreta (GK). Daneben werden 50 Prozent der Präfixe ⟨ver⟩ und ⟨vor⟩ (HM) korrekt verschriftet, aber Aufgaben zur Korrektur von Falschschreibungen hauptsächlich nicht bewältigt (Bremerich-Vos et al., 2012, S. 67 f.).

Kompetenzstufe II Die Verschriftung der elementaren Phonem-Graphem-Korrespondenzen erfolgt sicher und 75 Prozent der Schreibungen der Kategorie SG werden beherrscht. Auf orthografischer Stufe werden richtige Schreibungen zu 60 Prozent aus der Kategorie VK und zu 50 Prozent aus den Kategorien VL- und VA geleistet. Zudem wird von geglückten Schreibungen aus den Kategorienbereichen KA (in einfachen und strukturell komplexen Wörtern), HM (fast alle Wörter mit dem Suffix ⟨lich⟩ und 75 Prozent der Wörter mit den Präfixen ⟨ver⟩ und ⟨vor⟩), MG sowie GK (Substantive mit gegenständlicher Bedeutung und vereinzelt Abstrakta) gesprochen. Bei der Korrekturaufgabe können neben alphabetischen Schreibungen alle falschen Kleinschreibungen identifiziert werden (GK). Zudem gelingen Korrekturen im Bereich der Auslautverhärtung (KA) sowie der Interpunktion (Bremerich-Vos et al., 2012, S. 68 f.).

Kompetenzstufe III 80 Prozent der Testwörter werden korrekt geschrieben. Durchgängig gelingen die Großschreibung von Konkreta und Abstrakta, die als Substantive markiert sind, sowie die Schreibung aller Arten von Suffixen (HM). Ferner werden fast alle konsonantischen und vokalischen Ableitungen (KA und VA) und die Schreibung von ⟨qu⟩ (SG) beherrscht. Die Vokalkürze (VK) und die Markierung der Vokallänge mittels Dehnungs-h (VL-) gelingen zu 80 Prozent, sowie die Schreibung von Varianten des stimmlosen [s] (KA) erstmals in größerem Umfang. Die Korrekturleistung unterscheidet sich nur marginal von der auf Kompetenzstufe II. Allerdings werden Begründungen für Schreibung mit Auslautverhärtung angegeben (Bremerich-Vos et al., 2012, S. 69).

Kompetenzstufe IV Mehr als 90 Prozent der Testwörter werden korrekt geschrieben. Vollständig beherrscht wird die Kategorie spezieller Grapheme (SG), die Markierung der Vokalkürze an schwierigen bzw. komplexen Wortstellen (VK) sowie fast vollständig vokalische und konsonantische Ableitungen (VA, KA). Unbekanntere Wörter mit Dehnungs-h (VL-) werden teilweise korrekt verschriftet und erstmals Nominalisierungen richtig großgeschrieben (GK). Bei den Korrekturaufgaben werden Fehler bei der Vokalkürze (VK) und der vokalischen Ableitung (VA) aufgedeckt und können begründet werden. Ebenfalls gelingen Korrekturen im Hinblick auf das stimmlose [s] (Bremerich-Vos et al., 2012, S. 69 f.).

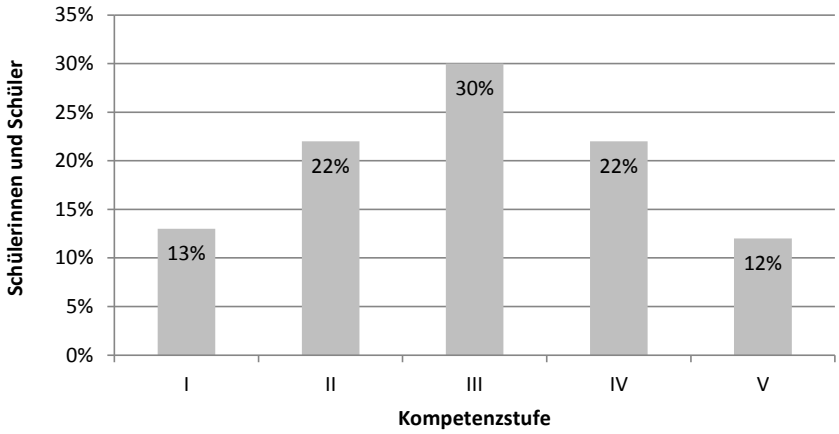



Abbildung 2.2: Verteilung der Schülerinnen und Schüler auf die Kompetenzstufen (nach Kultusministerkonferenz, Universität Duisburg-Essen & Institut zur Qualitätsentwicklung im Bildungswesen, 2013, S. 15)

Kompetenzstufe V Auf dieser Kompetenzstufe geht es um Wortschreibungen, für die syntaktisches Wissen notwendig ist, und mit schwer erkennbarer morphologischer Struktur. So werden z. B. nominalisierte Verben korrekt groß geschrieben (GK). Es werden alle Korrekturaufgaben in Form von Wahr-Falsch-Entscheidungen bewältigt (Bremerich-Vos et al., 2012, S. 70).

Die Verteilung der Schülerinnen und Schüler auf die Kompetenzstufen I bis V ist in Abbildung 2.2 dargestellt (durch Rundungen weicht die Summe der Prozentzahlen leicht von 100 ab). Hier zeigt sich, dass 13 Prozent der Kinder die Mindeststandards verfehlen, also am Ende von Klasse 4 und vor dem Eintritt in die Sekundarstufe I nicht die elementaren Laut-Buchstaben-Zuordnungen beherrschen. Sie bleiben somit deutlich hinter den von der KMK formulierten Erwartungen zurück. Auf der Stufe der Mindeststandards verharren 22 Prozent der Schülerinnen und Schüler. Insgesamt verfehlen damit 35 Prozent der Viertklässler die Regelstandards. Leistungen auf den Stufen der Regelstandards und Regelstandards plus erreichen 52 Prozent der Kinder. Kompetenzen, die oberhalb der Regelstandards liegen, und damit die rechtschriftlichen Erwartungen in der vierten Klasse übertreffen, erzielen 12 Prozent der Grundschülerinnen und Grundschüler.

### 2.1.4 Zusammenschau

Die Konzepte und die Validierung der beiden Rechtschreibtests HSP und AFRA wurden in diesem Abschnitt beschrieben. Für die HSP kann von einer guten Überprüfung des Strategiekonzepts ausgegangen werden, wobei May allerdings auf genauere Angaben bei der Darstellung der Modellüberprüfung verzichtet. So fehlen beispielsweise konkrete Werte zu der Gesamtmodellgüte oder inferenzstatistische Vergleichstests konkurrierender Modellvarianten. Für die AFRA-Fehlersystematik konnte keine empirische Evidenz der Separierbarkeit in die Fehlerkategorien erbracht werden. Die Auswertungen im Länder-

vergleich 2011 erfolgten als Konsequenz auf eindimensionaler Ebene mit differenzierten Fehlerkategorien in Anlehnung an die AFRA-Fehlersystematik. Dieses Modell diente auch als Vorlage für die dort definierten Kompetenzstufen.

In den beiden folgenden Abschnitten werden zwei weitere Rechtschreibtests mit unterschiedlich differenzierten Teilkompetenzen dargestellt. Hier zeigt sich einmal mehr, dass die beschriebenen Strategien, Fehlerkategorien, Prinzipien etc. zur Operationalisierung von Rechtschreibkompetenz heterogen sind, während sich in IGLU und PISA starke Überschneidungsbereiche in der Lesekompetenzermittlung ergeben (Bos et al., 2003, S. 84).

Durch den Einsatz verschiedener Rechtschreibtests mit ihren unterschiedlichen theoretischen Hintergründen sind die Leistungsergebnisse in der Folge studienübergreifend nicht 1:1 übertrag- und vergleichbar. Dennoch kann für den Bereich der Rechtschreibung eine Problemlage konstatiert werden. So wird studienübergreifend – wie u. a. hier für KESS und den IQB-Ländervergleich dargestellt – eine nicht unerhebliche Gruppe von rechtschreibschwachen Schülerinnen und Schülern ausfindig gemacht, die nicht die Kenntnisse der deutschen Orthografie aufweisen, die sie für ihre weitere schulische bzw. persönliche Laufbahn benötigen. Damit sind sie für den weiteren Lernerfolg (Deutschnote, Übergangsentscheidungen, Einflüsse auf andere Fächer, in denen Orthografie eine Rolle spielt) voraussichtlich negativen Konsequenzen ausgesetzt. Es zeichen sich damit Diskrepanzen zwischen den Zielen und Ansprüchen des Deutschunterrichts einerseits und den tatsächlich erreichten Kompetenzen andererseits ab.

## 2.2 gutschrift-diagnose

gutschrift-diagnose ist ein Test und Auswertungsverfahren von Ilona Löffler und Ursula Meyer-Schepers zur Diagnostik orthografischer Kompetenz. Vormals trug er den Namen Dortmunder Schriftkompetenzermittlung (DoSE).[14] Unter dieser Bezeichnung wurde der Test u. a. in IGLU-E 2001 eingesetzt (vgl. Abschnitt 2.2.2). Im Rahmen dieser Arbeit soll der Test bzw. das Konzept unter der aktuellen Bezeichnung gutschrift-diagnose bzw. gutschrift-Kompetenzmodell und auch kurz als gutschrift verwendet werden.

### 2.2.1 Theoretische Konzeption

gutschrift basiert auf der *Dortmunder Rechtschreibfehler-Analyse* (DoRA) (Valtin, Badel et al., 2003, S. 233). Daher soll zunächst eine kurze Beschreibung dieses Konzepts erfolgen, bevor darauf aufbauend die Erweiterungen des gutschrift-Rechtschreibkompetenzmodells erläutert werden.

[14] http://www.dose-diagnostik.de/Diagnostik.html, Zugriff am 13.08.2012

### DoRA als Vorläufer

DoRA ist ein Instrument zur Förderdiagnostik, das Verstöße gegen die Rechtschreibnorm qualitativ analysiert (Löffler & Meyer-Schepers, 1992, S. 7; Meyer-Schepers, 1991, S. 137). In den Jahren 1983 bis 1990 wurde DoRA im Institut für Legastheniker-Therapie und deutsche Orthographie (nun gutschrift-Institut zum Aufbau von Lese- und Schreibkompetenz) in Bochum und Dortmund von Löffler und Meyer-Schepers eingesetzt und auf Basis der theoretischen und praktischen Institutsarbeit weiterentwickelt (Löffler & Meyer-Schepers, 2008, S. 32; Löffler & Meyer-Schepers, 1992, S. 7). DoRA ist (wie auch die AFRA) eine orthografietheoretische Fehleranalyse, in der Fehler typologisiert werden, und wird daher auch als Fehlertypologie bezeichnet (Valtin, Badel et al., 2003, S. 229; Löffler & Meyer-Schepers, 1992, S. 15). Fehler werden dabei als Notwendigkeit angesehen, um den Schreibprozess zu analysieren und darauf basierend Fördermaßnahmen anzusetzen, die eine Fehlerminderung bezwecken (Meyer-Schepers, 1991, S. 137 f.). Falsche Schreibungen gelten nicht als Zufallsprodukte, sondern bringen „den Zusammenbruch der aktuellen Kompetenz des Schreibers in der Schreibsituation zur Anschauung“ (Meyer-Schepers, 1991, S. 221).

In DoRA werden Art, Ort und Umfang fehlerhafter Schreibungen ermittelt, voneinander unterschieden und differenzierten Fehlerquellen zugeordnet (Löffler & Meyer-Schepers, 1992, S. 15; Meyer-Schepers, 1991, S. 140). Über die Fehlerquellen werden auf einer ersten Ordnungsebene[15] Kategorien definiert. Folgende Kategoriengruppen werden von DoRA unterschieden: 1. Phonemfehler, 2. Graphemanordnung, 3. Dehnung/Dopplung, 4. Ableitung, Groß- und Kleinschreibung, Zusammen- und Getrenntschreibung und Silbentrennung sowie 5. Spezielle Phonem-Graphem-Zuordnungen und Sonderschreibweisen. Innerhalb der vierten Kategorie werden Schreibungen verortet, die eine „Anwendung des morphematischen, lexikalischen und syllabischen Prinzips“ erfordern (Löffler & Meyer-Schepers, 1992, S. 13). Hier werden also sowohl diverse rechtschriftliche Merkmale zu einer Kategorie zusammengefasst als auch viele Zugänge zu Wortschreibungen aufgenommen, deren Kenntnis mindestens auch teilweise in Kategorie 3 erforderlich ist.

Rechtschreibfehler sollen bei dem Konzept möglichst genau erfasst werden, um die Fehlerquellen festzustellen und in diesem Zuge die Schwierigkeiten einzugrenzen (Meyer-Schepers, 1991, S. 140). Sie werden den oben genannten Kategorien zugeordnet, welche die Basis für die richtigen Schreibungen bilden (Löffler & Meyer-Schepers, 1992, S. 14). Löffler und Meyer-Schepers (1992, S. 14) sprechen in diesem Zusammenhang von einer objektiven Unterscheidung und Definition der Fehler, da sie nicht über den Prozess des Zustandekommens – der immer subjektiv ist, da er unterschiedliche systematische oder zufällige Gründe haben kann – qualifiziert und katalogisiert werden. Entstehungsursachen von Falschschreibungen (wie z. B. mangelnde Fähigkeiten und Kenntnisse in Graphemformen, Phonemsegmentierung, grammatischen Regeln etc.) werden nicht direkt über die Kategorien benannt bzw. abgeleitet. Distanz wird zudem von deskriptiven Fehlerkatego-

[15] Eine ausführliche Beschreibung der Kategorien über die Zuordnung von Falschschreibungen liegt bei Meyer-Schepers (1991, S. 143 ff.) sowie Löffler und Meyer-Schepers (1992, S. 25 ff.) vor.

risierungen genommen, die ausschließlich mögliche Falschschreibungen benennen und aufsummieren: „Die Fehleranalyse ist als Diagnostikum jedoch wenig hilfreich, wenn sie rein deskriptiv im Sinne einer begriffslosen, d. h. von den schriftsprachlichen Teilleistungen aus gesehen unsystematischen Benennung der Falschschreibungen ausgeht.“ (Meyer-Schepers, 1991, S. 138, 140)

Die individuellen Fehler einer Person, die auf Grundlage eines Rechtschreibtests oder freier Schreibproben erhoben worden sind, werden über ein Fehlerraster ausgewertet. Das Raster bildet sich aus den oben genannten fünf Kategorien. Die einzelnen Fehler werden darin zur Erstellung individueller Fehlerprofile übertragen. Das Raster soll individuelle Fehlerschwerpunkte beschreiben und die Lernstände nach dem Konzept der DoRA verorten (Meyer-Schepers, 1991, S. 221).

### Das gutschrift-Rechtschreibkompetenzmodell

gutschrift übernimmt die Fehlerkategorien von DoRA. Eine Erweiterung stellt die Definition von differenziellen Rechtschreibkompetenzen dar: „Das mit DoSE zugrunde gelegte Modell der Orthographiekompetenz, das zugleich ein Fehlerkategoriensystem ist, möchte mit didaktischer Zielsetzung systematisch die Teilkompetenzen des Rechtschreiberwerbs kategorisieren.“ (Löffler & Meyer-Schepers, 2005, S. 82) Das Modell unterscheidet eine *phonographische* und eine *grammatische Kompetenzdimension* voneinander (Löffler & Meyer-Schepers, 2009, S. 61). Die in DoRA definierten Fehlerkategorien dienen dabei als Indikatoren für diese beiden Fähigkeiten. Die Entwicklung der beiden Kompetenzdimensionen erfolgt nach den Testautorinnen synchron (Valtin et al., 2004a, S. 142; Valtin, Badel et al., 2003, S. 234). „Zugleich will das Kompetenzmodell als ein *integratives Schriftsprachmodell* verstanden wissen, da es Schriftkompetenz nur in der Beschreibung beider Kompetenzdimensionen umfassend eingebettet sieht und deren Aufbau nicht in einem didaktischen Nacheinander, sondern in parallel verlaufenden Lernprozessen.“ (Löffler & Meyer-Schepers, 2007, S. 182) Trotz der theoretischen Differenzierung der zwei Zugriffsweisen werden häufig beide Fähigkeiten für die richtige Schreibung eines Wortes benötigt. Sie müssen vom Schreiber gleichzeitig angewendet werden, da die grammatische Dimension die phonographische Zugriffsweise ergänzt und korrigiert (Löffler & Meyer-Schepers, 2005, S. 84). „So setzt die grammatische Kompetenz dort ein, wo die lautanalytische Kompetenz nicht zur Orthographie verhilft (z. B. Morphemkonstanz, Formgestalt der jeweiligen Wortart).“ (Valtin, Badel et al., 2003, S. 229) Von „Entwicklungsmodellen“, die ein „Nacheinander von sich ablösenden Entwicklungsstufen mit jeweils dominanten Strategien“ beschreiben, distanzieren sich die Testautorinnen (Voss, Löffler, Meyer-Schepers, Meckel & Kowalski, 2008, S. 134). Die Vermittlung im Unterricht sollte demnach von beiden Fähigkeiten parallel erfolgen (Löffler & Meyer-Schepers, 2005, S. 84).

Die phonographische und grammatische Kompetenzdimension können wiederum die zwei Ausprägungen *elementare Stufe* und *erweiterte Stufe* annehmen, die sich kumulativ aufbauen (Löffler & Meyer-Schepers, 2009, S. 61; Voss, Löffler et al., 2008, S. 134; Löffler & Meyer-Schepers, 2008, S. 30; Löffler & Meyer-Schepers, 2007, S. 182). Diese

| | Fähigkeiten | |
|---|---|---|
| Niveaus | Elementar phonographisch | Elementar grammatisch |
| | Erweitert phonographisch | Erweitert grammatisch |

Tabelle 2.4: Das Rechtschreibkompetenzmodell von gutschrift

Aufgliederung soll die von Schriftlernenden zu leistenden einfachen bis schwierigen Schreiboperationen berücksichtigen (Voss, Löffler et al., 2008, S. 135; Valtin, Badel et al., 2003, S. 234). Löffler und Meyer-Schepers beziehen die Niveaus auf den Unterricht, da Schülerinnen und Schüler die Rechtschreibung vorwiegend im schulischen Kontext erlernen (Voss, Löffler et al., 2008, S. 134; Valtin et al., 2004a, S. 142). Die elementaren Anforderungen der phonographischen und grammatischen Kompetenzdimension sollten in den ersten beiden Klassenstufen beherrscht werden. Die Aneignung der erweiterten Kompetenzen sollte in den Klassenstufen 3 und 4 geschehen, wurde aber, auf Basis der Ergebnisse aus IGLU-E (vgl. Abschnitt 2.2.2), auf die folgenden zwei Jahrgangsstufen ausgeweitet (Voss, Löffler et al., 2008, S. 134; Valtin, Badel et al., 2003, S. 259).

Aus den differenzierten Kompetenzdimensionen und -stufen ergibt sich das in Tabelle 2.4 dargestellte gutschrift-Kompetenzmodell mit den vier Teilkompetenzen *elementar phonographisch*, *elementar grammatisch*, *erweitert phonographisch* und *erweitert grammatisch*.[16] Es zeigen sich 2x2 Kombinationen; die Testautorinnen bezeichnen das Modell daher auch als „2-2-Modell der Schriftkompetenz“ (Voss, Löffler et al., 2008, S. 134; Löffler & Meyer-Schepers, 2008, S. 30).

### Wörter und Indikatoren des Tests

Die Auswahl der Testwörter richtete sich nach den Grundwortschätzen der Bundesländer, damit die Wortbedeutungen den Schülerinnen und Schülern bekannt und die Anforderungen dem Lernstand angemessen sind (Löffler & Meyer-Schepers, 2005, S. 86; Valtin et al., 2004a, S. 141 f.). Löffler und Meyer-Schepers beschreiben im Rahmen des Einsatzes von IGLU 2001 einen Abgleich der Testwörter mit drei Grund- und Übungswortschätzen. Dieser ergab für Bayern eine 78-prozentige, für Thüringen eine 71-prozentige und für Niedersachen eine 65-prozentige Überschneidung (Valtin, Badel et al., 2003, S. 234).[17] Weitere Angaben zur curricularen Validität, z. B. bezogen auf eine Beurteilung von Experten aus den Ministerien der Bundesländer, wie dies für die weiteren in IGLU erhobenen Fächer (z. B. Lesen oder Mathematik) durchgeführt wird, finden sich nicht in der Literatur.

[16] Die vier Teilkompetenzen des Modells werden im Kontext der Datenskalierung auch als Subskalen oder Kompetenzdimensionen bezeichnet.

[17] Bremerich-Vos glich diese Wörter mit Naumanns Orientierungswortschatz für die Klassen 1 bis 6 ab und ermittelte eine 40-prozentige Überschneidung (Bremerich-Vos, 2004, S. 95).

Die in den Testwörtern enthaltenen Indikatoren[18] zur Operationalisierung der Teilkompetenzen wurden in Orientierung an die beschriebenen Rechtschreibphänomene in den Lehrplänen der Bundesländer ausgewählt (Valtin et al., 2004a, S. 143; Valtin, Badel et al., 2003, S. 234, 236). Die Indikatoren der elementaren phonographischen und grammatischen Kompetenz basieren auf den Anforderungen, wie sie für den Rechtschreibunterricht in den ersten beiden Schuljahren von den Lehrplänen aufgeführt sind; die der erweiterten Kompetenzen auf die dort beschriebenen Anforderungen der Folgeschuljahre (Valtin et al., 2004a, S. 142; Valtin, Badel et al., 2003, S. 237 f.): „Die Unterscheidung von elementarer und erweiterter Kompetenz will die Stufenfolge des Rechtschreiberwerbs in der Grundschule, wie er in den Lehrplänen mehr oder weniger explizit formuliert ist (erste / zweite bzw. dritte Klasse, Vertiefung in der vierten Klasse an schwierigeren Wörtern), systematisieren“ (Löffler & Meyer-Schepers, 2005, S. 84 f.). Einige ausgewählte Rechtschreibphänomene werden von dem Autorenteam konkret in unterschiedlichen Quellen benannt und als Indikatoren den Teilkompetenzen zugeordnet. In Tabelle 2.5 findet sich auf dieser Basis, u. a. bezugnehmend auf die IGLU-Berichtsbände (Löffler & Meyer-Schepers, 2005; Valtin et al., 2004a; Valtin, Badel et al., 2003), eine Zusammenführung einschließlich einiger Beispielwörter, sofern sie in der Literatur ebenfalls aufgeführt waren.

Indikatoren der elementaren phonographischen Kompetenz sind Phonem-Graphem-Korrespondenzen einschließlich spezieller Modifikationen. Im Einzelnen werden von Löffler und Meyer-Schepers das Schreiben von Vokalen, Diphthongen und Konsonanten sowie im speziellen von Affrikaten (⟨z⟩, ⟨pf⟩, ⟨qu⟩), Velarnasalen (⟨ng⟩ und ⟨nk⟩), Explosivlauten (⟨b⟩, ⟨t⟩, ⟨k⟩), Vokalen in unbetonten Lautpositionen und silbischen Konsonantenhäufungen genannt. Erweiterungen der Phonem-Graphem-Korrespondenzen, wie das alternative Verschriften derselben Sprachlaute (z. B. ⟨ei⟩ und ⟨ai⟩) sowie die Wiedergabe mehrerer Grapheme für ein Phonem (z. B. ⟨sch⟩, ⟨st⟩ und ⟨sp⟩ für /ʃ/) zählen ebenso zu der elementaren phonographischen Kompetenz. (Löffler & Meyer-Schepers, 2009, S. 70; Voss, Löffler et al., 2008, S. 149; Löffler & Meyer-Schepers, 2008, S. 30; Löffler & Meyer-Schepers, 2007, S. 184; Löffler & Meyer-Schepers, 2005, S. 87; Valtin et al., 2004a, S. 143; Valtin, Badel et al., 2003, S. 237)

Einige Indikatoren erfordern unabhängig von ihrer Zuordnung zum Bereich der elementaren phonographischen Kompetenz eine eingehendere Betrachtung. Dazu zählt die Schreibung von ⟨ng⟩ und ⟨nk⟩ für den Verlarnasal /ŋ/, da bei zweiterem zusätzlich ein /k/ produziert wird (vgl. z. B. die Unterscheidung bei Spannung und Bank). Eine Präzisierung ist ebenfalls bei dem Frikativ /ʃ/ erforderlich: auch hier müsste eine Differenzierung zwischen ⟨st⟩ und ⟨sp⟩, denen jeweils zusätzlich ein weiterer Laut folgt, und ⟨sch⟩ stattfinden. Daneben wird das Digraph ⟨qu⟩ unter die Affrikaten gefasst, was untypisch ist (Eisenberg, 2006c, S. 25), da der entsprechende Laut nicht homorgan ist. Der Artikulationsort der benachbarten Laute bei /pf/, /ts/ und /tʃ/ ist hingegen gleich oder sehr nahe gelegen (Altmann & Ziegenhain, 2007, S. 32, 38).

[18] Die Indikatoren werden im Verlauf der vorliegenden Arbeit auch als Analyseeinheiten und im Rahmen der Testauswertung als Items bezeichnet.

| Elementare phonographische Teilkompetenz | | Erweiterte phonographische Teilkompetenz | | Elementare grammatische Teilkompetenz | | Erweiterte grammatische Teilkompetenz | |
|---|---|---|---|---|---|---|---|
| **Indikatoren** | **Testwort-beispiele** | **Indikatoren** | **Testwort-beispiele** | **Indikatoren** | **Testwort-beispiele** | **Indikatoren** | **Testwort-beispiele** |
| Phonem-Graphem-Korrespondenzen | | Kürzezeichen in unflektierten Wortformen: <ck>, <mm>, <nn>, <tz>, <tt>, <ss> | Bä<c>ker, besti<m>men, spi<t>zen, Diske<t>te, aufgepa<s>st, verde<ck>t | Schreibung des F-Lauts<br>- als <v> in <ver>, <vor><br>- als <f> in <fer>, <for><br>- als <f> oder <v> | <v>erdeckt, <V>orsicht, in<f>ormieren, <f>ertigt, <V>ieh, <f>ließt | Kürzezeichen in flektierten Wörtern: <ck>, <mm>, <nn>, <tz>, <tt>, <ss> | aufgepa<ss>t, verde<c>kt |
| Vokale, Diphthonge (<ei>, <au>, <eu>), Konsonaten | | Dehnungs-e in unflektierten Wörtern | Vorspi<e>len | Vokalisierung des <r> | Bäch<er>, V<or>sicht | Dehnungs-e in flektierten Wörtern | zi<e>ht |
| Vokale in unbetonter Lautposition | Schreibm<a>schine, Kreuz<u>ng | Dehnungs-h | N<ah>rung, verkü<h>lt | Personalformen des Verbs | sink<t> | Ableitung des silbenanlautenden h | Vie<h>, dre<h>en, ru<h>ig, zie<h>t |
| Sonstige Konsonatenhäufungen (silbisch) | Mu<sk>eln, I<nt>eresse | Dehnungszeichen <aa> | | Kleinschreibung von Verb, Adjektiv, Adverb | spuken, spitzen, strömte, ölig | Ableitung des S-Lauts nach langem Vokal als <s> oder <ß> | flie<ß>t, schlie<ß>lich, verlo<s>t |
| Affrikaten<br>- <z><br>- <pf><br>- <qu> | glän<z>en, em<pf>indlich, über<qu>eren | | | Großschreibung konkreter Nomen und am Satzanfang | Bäcker, Muskeln | Großschreibung abstrakter Nomen und Nominalisierung | Interesse, Vorspielen, Kälte, Vorsicht |
| Velarnasal:<br>- <ng> und <nk><br>-Schreibung | Kreuzu<ng>, sti<nk>t | | | elementare Ableitung Verschlusslautung: <b>-<p>, <d>-<t>, <g>-<k> | Schrei<b>maschine | erweiterte Ableitung Verschlusslaute <b>-<p>, <d>-<t>, <g>-<k> | empfin<d>lich, Vorsicht<t> |
| Explosivlaute mit<br>- Folgevokal<br>- Folgekonsonant (unsilbisch) | <B>enzin<t>anks, ange<bl>ickt, <K>reuzung | | | elementare Endungen (Synkope in Wortendungen): <el>, <en>, <er>, <e> | hol<en>, Musk<el>n | Endsilben: <ig>, <lich>, <ung> | plötz<lich>, ruh<ig>, öl<ig> |
| alternative Verschriftung desselben Sprachlautes:<br>- z.B. <f>-<v>, <ai>-<ei>, <z>-<ts>, <ei>-<ai>, <au>-<ao><br>- <sch>, <st>, <sp> für den Sch-Laut (/ʃ/) | <schl>ießlich, be<st>immen, <Str>aße | | | elementare Vorsilben: <auf>, <an>, <be>, <ver>, <vor> | <ver>lost, <ver>brennen, <ver>kühlt, <Vor>spielen | Wortbildung (Zusammen-, Getrenntschreibung) | |
| | | | | Ableitung Umlaut:<br><ä> von <a>,<br><äu> von <au> | B<ä>cker, gl<ä>nzen, K<ä>lte, L<äu>fern | Deklinationsendungen | Spaziergang<s> |

Tabelle 2.5: Zuordnung von Indikatoren zu den gutschrift-Teilkompetenzen

Auf erweiterter phonographischer Ebene können Kinder in dem Kompetenzmodell Dehnungs- und Kürzezeichen zur Markierung der langen bzw. kurzen Akzentvokale setzen (Valtin et al., 2004a, S. 144; Valtin, Badel et al., 2003, S. 239). Darunter fallen die Dopplung des Konsonantengraphems (z. B. ⟨ck⟩, ⟨mm⟩) und die Schreibung des Dehnungs-e (⟨ie⟩) in unflektierten Wortformen sowie die Vokalgemination (⟨aa⟩) und das Dehnungs-h (Löffler & Meyer-Schepers, 2009, S. 70; Valtin et al., 2004a, S. 144; Valtin, Badel et al., 2003, S. 239).

Unter die grammatische Kompetenz werden Wortschreibungen gefasst, die von den einfachen und erweiterten Phonem-Graphem-Zuordnungen abweichen, sie korrigieren und ergänzen, da sie sich nicht über die Orientierung am Lautprinzip erschließen lassen, wie z. B. das Prinzip der Morphemkonstanz (Voss, Löffler et al., 2008, S. 135; Löffler & Meyer-Schepers, 2007, S. 184). Auf elementarer Stufe fallen darunter Ableitungen (Morphemkonstanz) von Umlauten (⟨ä⟩, ⟨äu⟩) und der Verschlusslaute bei Auslautverhärtung (⟨b⟩-⟨p⟩, ⟨d⟩-⟨t⟩, ⟨g⟩-⟨k⟩). Ebenso werden dazu wortgrammatische Grundkenntnisse gezählt, wie die Großschreibung am Satzanfang und von konkreten Nomen, die Kleinschreibung von Verben, Adjektiven und Adverbien sowie die Konjugation. Elementare Vorsilben und Endungen (z. B. ⟨en⟩, ⟨e⟩ und ⟨ver⟩, ⟨be⟩), die Vokalisierung des ⟨r⟩ sowie die alternative Verschriftung des F-Lauts (als ⟨f⟩ in ⟨fer⟩, ⟨for⟩ und als ⟨f⟩ oder ⟨v⟩, einschließlich des ⟨v⟩ in Vorsilben) sind weiterhin Indikatoren der elementaren grammatischen Teilkompetenz. (Löffler & Meyer-Schepers, 2009, S. 70; Voss, Löffler et al., 2008, S. 138 ff.; Löffler & Meyer-Schepers, 2008, S. 31; Löffler & Meyer-Schepers, 2005, S. 87 ff.; Valtin et al., 2004a, S. 143; Valtin, Badel et al., 2003, S. 237)

Die Verschriftung von Rechtschreibphänomenen in komplexen Wortformen (Flexion, Derivation und Komposita) bilden Analyseeinheiten der erweiterten grammatischen Kompetenz. Löffler und Meyer-Schepers nennen hier die erweiterte Ableitung von Verschlusslauten, die Schreibung von ⟨s⟩ oder ⟨ß⟩ nach langem Akzentvokal sowie von Deklinationsendungen (Löffler & Meyer-Schepers, 2007, S. 184; Valtin et al., 2004a, S. 144; Valtin, Badel et al., 2003, S. 240). Die Schreibung von Endsilben (z. B. ⟨ig⟩, ⟨ung⟩), die satzinterne Großschreibung abstrakter Nomen und Substantivierungen sowie der Bereich der Getrennt- und Zusammenschreibung werden von Kindern auf dieser Kompetenzstufe ebenfalls beherrscht (Löffler & Meyer-Schepers, 2009, S. 70; Löffler & Meyer-Schepers, 2005, S. 94; Valtin et al., 2004a, S. 144, 149; Valtin, Badel et al., 2003, S. 240). Als weitere Indikatoren werden die Konsonantengraphemverdopplung und das Dehnungs-e in flektierten Wörtern aufgeführt sowie die „Ableitung [...] des ‚stummen' -h, das erst silbenanlautend zum ‚Laut-h' wird (z. B. ‚ruig' statt ‚ruhig')"[19] (Löffler & Meyer-Schepers, 2005, S. 95). Im Zusammenhang mit der Verschriftung von Kürzezeichen werden von den Testautorinnen die Wörter aufgepasst und verdeckt aufgeführt (Valtin et al., 2004a, S. . 144). Hierbei merkt Bremerich-Vos (2004, S. 97) kritisch an, dass diese keine flektierten Wortformen sind. Zudem fällt eine inkonsistente Zuordnung bei den beiden Testwörtern bzw. den Analyseeinheiten auf. In den ersten beiden IGLU-Berichtsbänden werden sie der erweiterten

[19] Eine Auseinandersetzung mit dem silbenanlautenden h und weiteren ausgewählten Analyseeinheiten sowie deren Zuordnungen zu den Teilkompetenzen erfolgt in Abschnitt 2.4.

grammatischen, in dem dritten IGLU-Band hingegen der erweiterten phonographischen Teilkompezenz zugewiesen (Löffler & Meyer-Schepers, 2005, S. 91 f.; Valtin et al., 2004a, S. 144; Valtin, Badel et al., 2003, S. 240).

Indikatoren aus allen vier Teilkompetenzen sind beispielswese in dem Testwort Vieh vorhanden. Löffler und Meyer-Schepers (2007, S. 187) rechnen die Richtigschreibung des Wortes den folgenden orthografischen Anforderungsbereichen zu:

- Elementar grammatisch: Großschreibung des Anfangsbuchstabens als Großschreibung von Nomen
- Elementar grammatisch: ⟨v⟩-Schreibung nach der Regel alternativer Graphemzuordnungen (⟨v⟩ statt ⟨f⟩) [20]
- Elementar phonographisch: ⟨i⟩-Schreibung nach der Regel der elementaren Laut-Buchstaben-Zuordnung
- Erweitert phonographisch: ⟨e⟩-Schreibung (Dehnungs-e) nach dem lang zu sprechenden ⟨i⟩ zur Kennzeichnung der Länge des Akzentvokals
- Erweitert grammatisch: ⟨h⟩-Schreibung als Merkmal der Verschriftung des Silben-h

Bei der Schreibung von Vieh können demnach null bis fünf Einzelfehler auftreten.

### Fehleranalytische Auswertung

Löffler und Meyer-Schepers (2005, S. 85) stellen bei der Konzeption von gutschrift die Wichtigkeit heraus, alle Fehlerarten so vollständig und präzise wie möglich zu erfassen: „Die Aussagekraft einer groben Zuordnung von Fehlern zu einem der großen sogenannten Strategiebereiche, deren Abgrenzung im Übrigen variiert, ist gering“. gutschrift wird daher auch als eine Art systematisches Verzeichnis aller Fehlerquellen betrachtet (Löffler & Meyer-Schepers, 2005, S. 84). In IGLU-E 2001 wurden von den Testautorinnen beispielsweise 120.123 Einzelfehler von 3.391 Schülerinnen und Schülern, die 45 Wörter verschriften sollten, computergestützt analysiert (Löffler & Meyer-Schepers, 2005, S. 81). Fehlerquellen sind, so das Autorenteam, im Vergleich zu der Ursache eines Fehlers, nicht subjektiv (wie bereits oben im Zusammenhang mit DoRA erwähnt) (Löffler & Meyer-Schepers, 2005, S. 85). Nachdem ein Fehler dem Fehlerkategoriensystem zugeführt wurde, wird er linguistisch analysiert und einer Fähigkeit und Stufe zugeordnet (Löffler & Meyer-Schepers, 2009, S. 65; Voss, Löffler et al., 2008, S. 136). Für jede Verschriftungsvorschrift gibt es eine Zuordnung in die Teilkompetenzen: „Wir kategorisieren in dem Modell fehleranalytisch alle Zugriffsweisen, die der Rechtschreiblerner nutzen kann, um den getätigten Fehler zu vermeiden und zu korrekten Schreibprodukten zu gelangen.“

[20] In Lischeid (2006, S. 227) wird die Schreibung von ⟨v⟩ in Vieh der elementaren phonographischen Teilkompetenz zugeordnet. In IGLU 2001 ist die Schreibung des F-Lauts als ⟨v⟩ in Vieh ein Indikator der elementaren grammatischen Kompetenz (Valtin, Badel et al., 2003, S. 238). Der Zuordnung aus IGLU wird hier gefolgt, wobei beide nicht plausibel erscheinen. Diese Einschätzung wird in Abschnitt 2.4.2 näher erläutert.

Testwort: Elefanten
Schreibprodukt: *älewanttn

| Teilkompetenz | Fehleranalyse |
|---|---|
| Elementar phonographisch | Schreibung <w> statt <f> |
| Elementar grammatisch | fehlende Großschreibung |
| | Falschschreibung des Suffixes <en> |
| Erweitert phonographisch | falsche Verdopplung von <t> |
| Erweitert grammatisch | <ä> für unbetontes e |

Tabelle 2.6: Beispiel für eine qualitative Fehlerauswertung in gutschrift (in Anlehnung an Voss, Löffler, Meyer-Schepers, Meckel & Kowalski, 2008, S. 136)

(Löffler & Meyer-Schepers, 2005, S. 84) Die Fehler besitzen damit nicht die gleiche Wichtigkeit, sondern sind abhängig von der jeweiligen Teilkompetenz, und weisen damit unterschiedliche Qualitäten auf (Löffler & Meyer-Schepers, 2009, S. 64 f.). Ein Beispiel für eine linguistische Klassifizierung der Falschschreibung des Testwortes Elefanten kann Tabelle 2.6 entnommen werden.

### 2.2.2 Studienergebnisse

gutschrift-diagnose wurde bisher in den folgenden nationalen und regionalen Studien eingesetzt: IGLU-E 2001, IGLU Belgien, KESS 4, ELEMENT 4, LEO[21] und in der Wissenschaftlichen Begleitforschung zu *Frühdiagnose rechtschreibschwächerer Schülerinnen und Schüler*.[22] In diesem Abschnitt sollen die Studien beschrieben werden, die sich, wie ZuRecht, ebenfalls auf die vierte Jahrgangsstufe beziehen: IGLU-E 2001, IGLU-Belgien, KESS 4 und ELEMENT 4. Sie dokumentieren alle den Anteil an Wortfehlern, wie u. a. bei der Studie KESS, deren Befunde bereits in Abschnitt 2.1.1 beschrieben wurden. In den beiden IGLU-Erhebungen werden Ergebnisse zusätzlich differenziert zu den Teilkompetenzen ausgegeben, Fehlerhäufungen ausgewählter Indikatoren betrachtet sowie das Abschneiden von verschiedenen Leistungsgruppen gegenübergestellt.

Neben den Analysen zur Rechtschreibkompetenz der Kinder werden zudem die Analysen zur Kompetenzmodellierung und -überprüfung aus der wissenschaftlichen Begleitfor-

[21] Die *Lernstands-Ermittlung und -Förderung schulischer Orthografiekompetenz* ist eine Bochumer Untersuchung in Klassenstufe 5. An ihr beteiligten sich über 1.000 Schülerinnen und Schüler im Jahr 2005 (Lischeid, 2007; Lischeid, 2006).

[22] Auf der Homepage des Löffler-Instituts wird ebenfalls das Projekt *Individuelle Förderung an Berufskollegs in NRW* erwähnt. Zu diesem Projekt liegen jedoch bisher noch keine Veröffentlichungen vor (http://www.loeffler-institut.de/schriftkompetenz-berufskollegs, Zugriff am 08.04.2014).

schung „Frühdiagnose rechtschreibschwächerer Schülerinnen und Schüler“ betrachtet, da hier erstmals die beiden elementaren Teilkompetenzen validiert worden sind. Eine vollständige Überprüfung des Kompetenzmodells – d. h. einer voll ausgebildeten Rechtschreibkompetenz – blieb bis zu diesem Zeitpunkt aus. Die vorliegende Arbeit schließt damit dieses Forschungsdesiderat. Die Befunde zum Kompetenzstand und -zuwachs der Kinder sind dabei nicht dokumentiert, da sie auf die zweite Klassenstufe bezogen sind und sich damit nicht als Referenz für die Ergebnisse aus ZuRecht anbieten. Dieser Abschnitt ergänzt damit den in Abschnitt 2.1 berichteten Forschungsstand.

### IGLU 2001

DoSE wurde im Rahmen der nationalen Ergänzungsstudie von IGLU 2001 in 12 Bundesländern (nicht teilgenommen haben Brandenburg, Mecklenburg-Vorpommern, Niedersachsen und Sachsen-Anhalt) eingesetzt. Insgesamt haben 3.615 Schülerinnen und Schüler aus 153 Schulen den Test innerhalb eines rotierten Designs bearbeitet (Valtin, Badel et al., 2003, S. 235). Er umfasste 45 Testwörter, die in vorgegebene Lücken von 19 Sätzen einzutragen waren (Valtin, Badel et al., 2003, S. 233). Die Diktierzeit betrug ca. 20 Minuten (Valtin, Badel et al., 2003, S. 234).

*Ergebnisse auf Ganzwort- und Teilkompetenzebene* Die quantitativen Analysen auf Ganzwortebene basieren auf einer Stichprobe von 2.951 Kindern und umfassen Ergebnisse zur Anzahl richtig geschriebener Wörter und zur Fehlerdichte (Valtin, Badel et al., 2003, S. 235). Den Test haben drei Schülerinnen und Schüler vollständig ohne Fehler bearbeitet. Durchschnittlich wurden von den 45 Diktatwörtern 25,6 Wörter vollständig richtig geschrieben, bei einer Standardabweichung von 9,0. Bei einer Gegenüberstellung der leistungsstärksten und -schwächsten fünf Prozent der Kinder in Bezug auf das Verhältnis von Wort- und Einzelfehlern zeigt sich eine große Leistungsspreizung zwischen diesen beiden Gruppen. So erzeugen die stärkeren fünf Prozent der Schreiber maximal 5 Fehler und die schwächeren fünf Prozent mehr als 63 Fehler bei der Schreibung der 45 Wörter (Valtin, Badel et al., 2003, S. 236).

Neben den Informationen zur Verschriftung vollständiger Wörter wird das durchschnittliche Abschneiden in den Teilkompetenzen dokumentiert. Die Indikatoren der elementaren grammatischen Kompetenz werden am häufigsten fehlerfrei verschriftet: der Durchschnitt liegt bei 92,8 Prozent Richtigschreibungen. Darauf folgt der Anteil an Richtigschreibungen in der elementaren phonographischen Subskala mit 88,3 Prozent und in der erweiterten grammatischen Teilkompetenz mit 70,8 Prozent. Am schwierigsten fällt den Schülerinnen und Schülern die Beherrschung der erweiterten phonographischen Teilkompetenz, bei der 66,9 Prozent der Indikatoren richtig geschrieben werden. (Valtin, Badel et al., 2003, S. 241)

*Schwierigkeitsgrade ausgewählter Indikatoren* Erweitert werden die Angaben zu den Teilkompetenzen in IGLU um die Betrachtung einzelner ausgewählter Indikatoren. Die Analysen beziehen sich auf 3.391 Schülertests (Löffler & Meyer-Schepers, 2005, S. 81). Die Stichprobe wurde auf Basis ihrer Einzelfehler in drei Gruppen untereilt: in ein oberes Viertel (die oberen 25 Prozent, 847 Kinder), ein Mittelfeld (25-75 Prozent, 1.695 Kinder) und ein unteres Viertel (untere 25 Prozent, 849 Kinder) (Löffler & Meyer-Schepers, 2005, S. 82 f.). Die Auswertungen umfassen Ergebnisse zu den prozentualen Falschschreibungen von Analyseeinheiten, getrennt nach den drei Gruppen und zugeordnet zu den Teilkompetenzen. Die Darstellung der Befunde folgt den von dem Testautorinnenteam vorgenommenen Klassifizierungen der Indikatoren, auf deren spezifische Unschärfen bereits bei der Beschreibung der Teilkompetenzen eingegangen wurde; eine weiterführende kritische Auseinandersetzung findet in Abschnitt 2.4 statt.

Die stärksten Probleme innerhalb der elementaren phonographischen Kompetenz sind in allen drei Gruppen bei der Verschriftung der Velarnasale (⟨ng⟩, ⟨nk⟩) zu finden. Im unteren Viertel werden 34,0 Prozent, im Mittelfeld 17,5 Prozent und im oberen Viertel 5,1 Prozent der Indikatoren fasch geschrieben. Weiterhin zeigen sich Unsicherheiten bei den Affrikaten (als ⟨z⟩, ⟨pf⟩, ⟨qu⟩), deren Fehleranteil in den Gruppen 29,0 (unteres Viertel), 12,6 (Mittelfeld) und 3,8 Prozent (oberes Viertel) beträgt. Bei den weiteren Analyseeinheiten – Verschlusslaute, Schreibung von ⟨st⟩, ⟨sp⟩ sowie ⟨sch⟩ – zeigen sich in der oberen und mittleren Gruppe keine bis wenig Probleme mit maximal 2,5 Prozent Fehlern. Im unteren Viertel sind diese Bereiche deutlich fehlerträchtiger und variieren zwischen 6,3 und 8,5 Prozent. Löffler und Meyer-Schepers (2005, S. 86 f.) halten fest, dass die elementare phonographische Kompetenz bei den meisten Schülerinnen und Schülern gut ausgeprägt ist.

Im Bereich der elementaren grammatischen Teilkompetenz weisen Löffler und Meyer-Schepers die folgenden drei Fehlerschwerpunkte aus: Ableitung des Umlautes, Vorsilben (⟨ver⟩, ⟨vor⟩) sowie Groß- und Kleinschreibung konkreter Nomen, Adjektive und Verben. Im unteren Viertel werden durchschnittlich mehr als die Hälfte (55,1 Prozent) und im Mittelfeld rund ein Viertel (25,7 Prozent) der Ableitungsoperationen des Umlautes nicht bewältigt. Der Fehleranteil im oberen Viertel beträgt hier 5,5 Prozent. Er fällt mit 7,7 Prozent in dieser Gruppe am höchsten bei den Vorsilben aus. Für die beiden weiteren Gruppen ist dies der zweitgrößte Problembereich, mit 41,9 (unteres Viertel) und 21,3 (Mittelfeld) Prozent Falschschreibungen. Bei der Groß- bzw. Kleinschreibung werden von den leistungsschwächsten 25 Prozent der Kinder durchschnittlich 11,9 Prozent der Indikatoren falsch geschrieben. Im oberen Viertel und Mittelfeld gibt es hier keine bis wenige Probleme (0,0 Prozent und 3,6 Prozent). (Löffler & Meyer-Schepers, 2005, S. 90)

Bei der erweiterten phonographischen Kompetenz zeigen insbesondere die Schülerinnen und Schüler der mittleren und unteren Gruppe größere Unsicherheiten. Die Dehnungs- und Kürzezeichen werden von 68,1 Prozent des unteren Viertels, von 39,5 Prozent der Kinder im Mittelfeld sowie von 15,5 Prozent des oberen Viertels nicht richtig verschriftet. Der Fehler kann dabei unterschiedliche Ausprägungen haben: Dehnungs- und Kürzezeichen können in Kombination verwendet werden (*hohllen). Zudem können die phonematische

Länge bzw. Kürze des Akzentvokals (*spucken anstatt spuken, *bestiemen) sowie der phonematische Ort des Akzentvokals (*Behnzintanks, *Muskelln, *plötzliech) vertauscht werden (Löffler & Meyer-Schepers, 2005, S. 91 ff.). Löffler und Meyer-Schepers (2005, S. 93) bewerten diese Ergebnisse wie folgt: „Insgesamt kann von einer zufrieden stellenden Rechtschreibsicherheit in der Setzung von Dehnungs- und Kürzezeichen bei deutschen Viertklässlern nicht gesprochen werden."

Zu den Indikatorbereichen Großschreibung abstrakter Nomen und Nominalisierungen, Ableitung sowie Deklination liegen für die erweiterte grammatische Subskala Ergebnisse vor. Die größten Probleme zeigen sich hier bei der Großschreibung, bei der im unteren Viertel 46,1, im Mittelfeld 36,1 und im oberen Viertel 20,3 Prozent Fehler gemacht werden. Der Bereich der Ableitung wurde bei der Darstellung unterteilt und die Wortbildungssuffixe ⟨ig⟩, ⟨lich⟩ und ⟨ung⟩ separat ausgewiesen. Die leistungsschwächsten 25 Prozent der Kinder schreiben 33,8 Prozent der Ableitungsindikatoren (wie z. B. Ableitung der Auslautverhärtung oder des silbenanlautenden h) falsch und 14,4 Prozent der speziell ausgewiesenen Suffixe. Bei der mittleren Gruppe und den leistungsstärksten 25 Prozent der Kinder beträgt der Fehleranteil bei den Ableitungen 19,2 (Mittelfeld) und 5,9 (oberes Viertel) Prozent sowie bei den Suffixen 5,0 (Mittelfeld) und 0,0 (oberes Viertel) Prozent. Bei der Flexion weisen die Schülerinnen und Schüler aus dem oberen Viertel ebenfalls keine Unsicherheiten auf, und die Kinder aus dem Mittelfeld eine durchschnittliche Fehleranzahl, die 5 Prozent entspricht. Das untere Viertel schreibt dagegen 9,1 Prozent der Indikatoren falsch. (Löffler & Meyer-Schepers, 2005, S. 95)

Löffler und Meyer-Schepers resümieren, dass die Kinder des oberen Viertels lernstandstypische Fehler bei den erweiterten Teilkompetenzen machen, die voraussichtlich mit der Zeit weniger werden und verschwinden. Sie sprechen von einem guten orthografischen Entwicklungsstand dieser Schülergruppe auf Basis des geringen Anteils an Fehlern. Für die Schülerinnen und Schüler des unteren Viertels wird von dem Autorenteam kein vergleichbarer Lernstand ausgewiesen. Diese Gruppe zeigt starke Unsicherheiten, die sich in der Anzahl an Fehlern und Schreibvarianten, sowie an Fehlern aus allen Kompetenzbereichen erkennbar machen (Löffler & Meyer-Schepers, 2005, S. 96 ff., 101 ff.). So produziert das untere Quartil der Schülerinnen und Schüler im Durchschnitt 113 Variantenschreibungen pro Wort, während das obere Quartil im Mittel 21 Varianten aufweist. Löffler und Meyer-Schepers (2005, S. 96) demonstrieren dies u. a. anhand der Schreibungen des Wortes verkühlt, das im oberen Viertel in 10 und im unteren Viertel in 107 unterschiedlichen Varianten verschriftet wird. „Die im Unterricht erzielte Klärung, worin eine Schreibvorschrift besteht, welche Anwendungsbedingungen zu beachten sind und wie sie zu anderen orthographischen Vorschriften im Verhältnis steht, ist bei ihnen nicht nur nicht angekommen [. . . ] – in ihre Rechtschreibbemühungen ist mehr oder weniger stark Verunsicherung eingekehrt, ein ungelöstes Problem zieht weitere Probleme nach sich." (Löffler & Meyer-Schepers, 2005, S. 103) Ein genaueres Fehlerbild dieser Gruppe wird in IGLU 2005 beschrieben. Insgesamt kann festgehalten werden, dass von keiner ausgebildeten orthografischen Kompetenz aller Viertklässler und keinem homogenen Leistungsstand ausgegangen werden kann.

### IGLU Belgien

IGLU Belgien wurde an PIRLS 2006 angeschlossen und ist eine Erhebung in der deutschsprachigen Gemeinschaft von Belgien. Die wissenschaftliche Leitung der Studie trug Wilfried Bos (Bos, Sereni & Stubbe, 2008, S. 16). Neben der Domäne des Lesens und den Hintergrundmerkmalen wurde die Rechtschreibleistung erhoben. Die Stichprobe beträgt 866 Schülerinnen und Schüler der vierten Klasse, die ein Lückensatzdiktat mit 35 Testwörtern, die in 9 Sätze einzutragen waren, bearbeitet haben (14 der 35 Testwörter wurden bereits in IGLU eingesetzt). Vier Kinder schreiben alle Wörter des Lückentextes fehlerfrei. Durchschnittlich werden 14,3 Wörter falsch geschrieben, bei einer Standardabweichung von 7,5 (Löffler, Meyer-Schepers & Stubbe, 2008, S. 137 ff.).

Auf Basis der Wortfehler wurde die Stichprobe in Quartile eingeteilt. Kinder mit 0 bis 8 Fehlern werden dem 1. Quartil zugeordnet (24,5 Prozent der Stichprobe), mit 9 bis 13 Fehlern dem 2. Quartil (25,1 Prozent), mit 14 bis 19 Fehlern dem 3. Quartil (26,2 Prozent) und mit 20 oder mehr Fehlern dem 4. Quartil (24,2 Prozent). Das Abschneiden der Schülerinnen und Schüler in den Teilkompetenzen wird über die Beschreibung der Leistungsdifferenzen zwischen dem 1. und 4. Quartil dargestellt. Dabei werden allerdings keine prozentualen Angaben und keine Angaben zur jeweiligen Indikatorenanzahl der Teilkompetenzen dokumentiert. Die durchschnittliche Anzahl an Falschschreibungen, die das 1. Quartil in den Teilkompetenzen tätigt, liegt zwischen ca. 1 und 3 Fehlern. Bei den Schülerinnen und Schülern des 4. Quartils fällt die Anzahl deutlich höher aus. Es werden in der elementaren phonographischen Kompetenz ca. 7,5, in der elementaren grammatischen Kompetenz ca. 5, in der erweiterten phonographischen Kompetenz ca. 16 und in der erweiterten grammatischen Kompetenz ca. 15 Fehler gemacht. Die drei größten Fehlerquellen sind die Setzung von Dehnungs- und Kürzezeichen, Ableitungen sowie die Großschreibung. Im Mittel werden bei den Indikatoren des ersten Bereichs ca. 9 Fehler (gegenüber ca. 3 Fehlern im 1. Quartil) und des zweiten und dritten Bereichs ca. 7 Fehler (ca. 1 bis 2 Fehler im 1. Quartil) produziert (Löffler et al., 2008, S. 139 ff.). Löffler und Meyer-Schepers konstatieren, dass die Leistungsstände der Kinder weit auseinandergehen und Schriftlernende des 4. Quartils, wie bereits in IGLU 2001 beobachtet, Defizite in allen Teilkompetenzen aufweisen, und dementsprechend einer Förderung bedürfen (Löffler et al., 2008, S. 144 f.).

### ELEMENT

Für die Leitung der Studie ELEMENT (*Erhebung zum Lese- und Mathematikverständnis – Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin*) wurde von der Berliner Senatsverwaltung für Bildung, Jugend und Sport Rainer Lehmann beauftragt. ELEMENT wird in Berlin mit dem Ziel durchgeführt, die Lernausgangslage der Schülerinnen und Schüler und die Lernfortschritte zwischen der Primar- und Sekundarstufe zu untersuchen (Lehmann & Nikolova, 2005, S. 7). Es handelt sich also um eine Längsschnittstudie mit Erhebungszeitpunkten in 2003, 2004 und 2005. Der gutschrift-Test wurde ausschließlich

zur Eingangserhebung (ELEMENT 4) eingesetzt, während z. B. Lese- und Mathematikverständnis zu allen drei Messzeitpunkten erhoben worden sind und die Lernentwicklung im Abschlussbericht dargestellt ist (Lehmann & Lenkeit, 2008, S. 12, 17–38). Die Rechtschreibung wurde mit dem gleichen Testinstrument von gutschrift ermittelt, wie bereits zuvor in IGLU (Lehmann & Nikolova, 2005, S. 12).

Die Stichprobe bestand bei der Eingangserhebung aus allen Berliner Grundschulkindern, die im Schuljahr 2002/2003 die vierte Klassenstufe in staatlichen Schulen besuchten, und ist damit repräsentativ für den Stadtstaat (Lehmann & Nikolova, 2005, S. 9). Für gutschrift liegen Ergebnisse zu den Daten von 3.148 Schülerinnen und Schülern vor (Lehmann & Nikolova, 2005, S. 26). Sie werden ausschließlich nach dem Anteil der Wortfehler ausgewertet. Im Mittel wurden von den 45 Wörtern 23,2 vollständig richtig geschrieben. Die Standardabweichung beträgt 9,5. Der Mittelwert liegt damit knapp unter, und die Leistungsvariation knapp über den Werten in IGLU-E 2001. Die Ergebnisse weichen nur marginal von den Befunden aus KESS ab (vgl. Abschnitt 2.1.1).

#### Wissenschaftliche Begleitforschung zu Frühdiagnose rechtschreibschwächerer Schülerinnen und Schüler

Die wissenschaftliche Begleitforschung des gutschrift-Projekts zu Rechtschreibdiagnose und -förderung geschah unter Leitung von Wilfried Bos und Andreas Voss zusammen mit der Autorin dieser Arbeit. Bei dem Projekt handelt es sich um eine Kooperation zwischen dem gutschrift-Institut und dem Regionalen Bildungsbüro der Stadt Dortmund, mit dem Ziel, Versagen in Schriftspracherwerbsprozessen zu einem möglichst frühen Zeitpunkt zu erkennen und zu verhindern. Es wurde ein quasiexperimentelles Untersuchungsdesign mit einer Lernausgangslagen- und einer Folgeerhebung in Jahrgangsstufe 2 eingesetzt. Das Projekt wurde im Schuljahr 2006/2007 durchgeführt und es beteiligten sich 29 Grundschulen mit 1.578 Schülerinnen und Schülern aus Dortmund (Voss, Löffler et al., 2008, S. 123). Die Analysen basieren auf den Daten des gutschrift-1-Tests. Es handelt sich dabei um einen Test in Lückensatzformat mit 23 Sätzen und 23 Diktatwörtern, der für die ersten beiden Klassenstufen konzipiert ist (Voss, Löffler et al., 2008, S. 136; Löffler & Meyer-Schepers, 2007, S. 184). Mit ihm werden die elementaren, aber nicht die erweiterten Fähigkeiten der Schülerinnen und Schüler erfasst. Die Skalierung der Daten erfolgte demnach nur über einen Teil des gutschrift-Kompetenzmodells bzw. für ein Niveau (Voss, Löffler et al., 2008, S. 144). Eine vollumfängliche Analyse des gutschrift-diagnose-Modells findet erstmals im Rahmen dieser Arbeit statt.

Für die elementare phonographische und elementare grammatische Teilkompetenz wurden Korrelations- sowie Reliabilitätsanalysen im Rahmen des Projektes durchgeführt. Zudem erfolgte ein Vergleich des Modells mit einem alternativen Modell, das keine Dimensionen differenziert.[23] Die beiden fähigkeitsbasierten Dimensionen umfassten jeweils 29 Analyseeinheiten. Es wurde ein Zusammenhang von 0,58 zwischen den beiden Subskalen

[23] Details zu den statistischen Berechnungen und zur Einordnung der Werte finden sich in Abschnitt 4.2.1. Dort werden die hier berichteten Ergebnisse ebenfalls mit den Werten aus ZuRecht verglichen.

berechnet. Die Reliabilitäten nehmen Werte von 0,63 für die phonographische Kompetenz und 0,76 für die grammatische Kompetenz an. Die Gegenüberstellung eines eindimensionalen Modells mit dem zweidimensionalen Modell (in dem modellkonform die beiden Teilfähigkeiten unterschieden werden) zeigt eine signifikant bessere Datenanpassung in Bezug auf das differenzierte Kompetenzmodell (Voss, Löffler et al., 2008, S. 145 f.). Die Autoren folgern: „Die strukturelle Unterscheidung zwischen phonographischer und grammatischer Teilkompetenz ist aus analytischer Sicht sinnvoll." (Voss, Löffler et al., 2008, S. 151)

## 2.3 Sprachsystematischer Rechtschreibtest

Inge Blatt ist die Autorin des Sprachsystematischen Rechtschreibtests (SRT), der in Kooperation mit Andreas Voss entwickelt wurde. Der SRT wurde im Rahmen der Voruntersuchung zu IGLU-E 2006 das erste Mal eingesetzt, um die theoretische Konzeption empirisch zu prüfen (vgl. Abschnitt 2.3.2). Seitdem ist er weiterentwickelt und in verschiedenen Studien eingesetzt worden. Im Folgenden wird zunächst die linguistische Grundlage des SRT, auf der das Kompetenzmodell aufbaut, und im Anschluss der Einsatz und die Ergebnisse des Tests in den unterschiedlichen Studien beschrieben.

### 2.3.1 Theoretische Konzeption

Die Konzeption des SRT basiert auf den graphematischen Forschungsergebnissen von Peter Eisenberg und Gabriele Hinney (Blatt, Voss, Kowalski & Jarsinski, 2011, S. 236). Daher sollen diese nachfolgend dargestellt werden.

#### Graphematik nach Eisenberg als Grundlage

Die Graphematik untersucht das Schriftsystem einer Sprache, so wie die Phonologie das Lautsystem der Sprache untersucht. Als Vertreter der *Interdependenzthese* betrachtet Eisenberg das Schriftsystem und das Lautsystem als gleichberechtigte und eigenständige Teilsysteme der deutschen Sprache. In diesem Sinne ist das Geschriebene kein Abbild des Gesprochenen. Um die Beziehung des Schrift- und Lautsystems zueinander zu beschreiben, untersuchte Eisenberg zunächst beide Systeme separat voneinander und bezog sie im Anschluss aufeinander. Er zeigte Gemeinsamkeiten und Besonderheiten der Struktur der geschriebenen und der gesprochenen Sprache auf (Butt & Eisenberg, 1990; Eisenberg, 1989; Eisenberg, 1983). Als Ergebnis formulierte er die Annahme eines regelhaften Schriftsystems, das in einem systematischen Verhältnis zum Lautsystem steht. Danach lässt sich ein sogenannter *Kernbereich* (im Gegensatz zum *Peripheriebereich*, s. u.) der Rechtschreibung, der 90 bis 95 Prozent des nativen Wortschatzes umfasst, durch vier Prinzipien regeln: das phonographische, das silbische, das morphologische und das wortübergreifende Prinzip (Eisenberg & Fuhrhop, 2007, S. 24 f.; Eisenberg, 2006c, S. 61 ff.). „Im Deutschen weicht

die orthographische Norm lediglich in Kleinigkeiten und sehr beschränktem Umfang von einem graphematisch rekonstruierbaren Schriftsystem ab.“ (Eisenberg, 2006a, S. 304) Diese Prinzipien sollen nachfolgend erläutert werden, da der SRT diese beinhaltet. Dabei werden zum Teil Grundlagen des Aufbaus der deutschen Sprache kurz umrissen, wenn diese für die Erläuterung von Eisenbergs Ansatz erforderlich sind.

*Phonographisches Prinzip* Das phonographische Prinzip basiert auf Graphem-Phonem-Korrespondenzen bzw. Graphem-Phonem-Korrespondenzregeln (Eisenberg, 2006a, S. 309). Diese besagen, welches Graphem, bzw. welche Graphemfolge, welchem Phonem bzw. welcher Phonemfolge entsprechen. Über das phonographische Prinzip lässt sich beispielsweise die Schreibung des Wortes Schokolade erklären. Das geschriebene Wort bildet das gesprochene Wort ab und umgekehrt.[24]

Eine ausschließlich phonographische Schreibung reicht aber bei vielen Wörtern nicht aus, um sie zu differenzieren und orthografisch richtig zu schreiben (Hinney, 1997, S. 90). So ist z. B. die richtige Schreibung von Sonnenliege oder Abendruhe rein phonographisch nicht herleitbar. Ebenfalls bewirkt z. B. die Aussprache des Vokals in den Wörtern kann und Kahn eine semantische Unterscheidung, die einer rein phonographischen Schreibweise, wie z. B. *kan, nicht zu entnehmen wäre. Um eine korrekte Schreibung im Deutschen herbeizuführen, bedarf es der Beachtung „der Struktur größerer sprachlicher Einheiten, nämlich der Silbe, dem Morphem und der Wortform.“ (Eisenberg, 1988, S. 150)

*Silbisches Prinzip* Das silbische Prinzip ist Kernstück der Graphematik nach Eisenberg. Die „Schreibsilbe“ bildet innerhalb dieses Prinzips die Basis. Eisenberg führte diesen Begriff ein, um die gesprochene Silbe von der geschriebenen Silbe zu unterscheiden (Eisenberg, 1989, S. 60 ff.). Die Silbe wird in Anfangsrand, Kern und Endrand untergliedert. Kerne sind immer besetzt, bestehen aus einem Vokal oder einem Diphthong und bestimmen die Anzahl an Silben eines Wortes. Anfangs- und Endränder setzen sich aus Konsonanten zusammen. Aus genau einem Laut bestehende Ränder werden als einfache Ränder, aus einer Lauthäufung bestehende Ränder als komplexe Ränder bezeichnet (Eisenberg, 2006a, S. 100 ff.; Fuhrhop, 2006, S. 14). Das Wort Stein besteht z. B. aus einem komplexen Silbenanfangsrand und einem einfachen -endrand (sowie dem obligatorischen Kern) und das Wort froh aus einem komplexen Anfangs- und einem leeren Endrand. Der Anfangsrand ist bei wortanlautenden Silben aufgrund des Glottisverschlusses [ʔ], wie z. B. bei ernst, nicht wirklich leer (Altmann & Ziegenhain, 2007, S. 92; Maas, 2006, S. 217).

Voraussetzung für silbische Schreibungen und Regeln (s. u.) ist die Bestimmung von Silbengrenzen. Nach Eisenberg (2006c, S. 37 f.) ist dies eine Fähigkeit, die Kinder ohne Training beherrschen. „Die Gliederung von Wortformen in Silben ist dem Sprecher intuitiv zugänglich. Ohne Schwierigkeiten lässt sich angeben, wie viele Silben eine Wortform hat. Kinder verfügen über diese Kenntnis genauso wie Erwachsene. Bevor Kinder schreiben lernen, wissen sie im Allgemeinen nicht, dass Wortformen aus Lautsegmenten aufgebaut

[24] Die Graphem-Phonem-Korrespondenzregeln sind in Eisenberg (2006a, S. 307 ff.) dargestellt.

sind. Dagegen machen viele Kinderspiele von der Gliederung der lautlichen Formen in Silben Gebrauch (z. B. Abzählreime).“ (Eisenberg, 2006c, S. 37 f.) Die syllabische Struktur (Silbenanzahl und Lage der Silbengrenze) der Wörter übernimmt eine strukturierende Funktion, die für Schülerinnen und Schüler transparent ist, da sie in der phonologischen Wortform mit der graphematischen übereinstimmt (Butt & Eisenberg, 1990, S. 56 ff.). Im Folgenden werden die Besonderheiten der silbischen Schreibung bzw. das Verhältnis von Schreib- und Sprechsilbe im Anfangsrand, Kern und Endrand der Silbe nach Eisenberg vorgestellt.

Der Anfangsrand einer Schreibsilbe besteht aus maximal drei Konsonantengraphemen (Eisenberg, 2006a, S. 115). Er kann zumeist über das phonographische Prinzip erschlossen werden, so z. B. bei den Wörtern Schwester oder Glas (Eisenberg, 2006a, S. 311). Bei der Schreibung des Lauts [ʃ] vor [t] oder [p] findet eine Abweichung vom Phonographischen statt. Stehen [ʃ] und [t] oder [ʃ] und [p] gemeinsam im Anfangsrand, folgen zumeist weitere Konsonanten. Um eine Überlänge des Anfangsrandes zu vermeiden, findet, so Eisenberg, ein Längenausgleich statt, um beispielsweise Schreibungen mit fünf Konsonanten auszuschließen (*Schtrumpf, *Schprache) (Eisenberg, 2006c, S. 71; Butt & Eisenberg, 1990, S. 55; Eisenberg, 1989, S. 67). Eisenberg erläutert, dass silbische Einheiten ungefähr gleich große Graphemcluster aufweisen, wodurch die silbischen Informationen dem Leser schneller zugänglich sind. Die Schrift strebt nach einer Ausgewogenheit in der Silbenlänge (Eisenberg, 2006a, S. 311).

In der Phonologie werden 16 Vokale voneinander unterschieden, die neun Vokalgraphemen in der Graphematik entsprechen (Eisenberg, 2006c, S. 67; Fuhrhop, 2006, S. 9; Butt & Eisenberg, 1990, S. 55). Da dies deutlich weniger sind, müssen mehrere Phoneme über ein Graphem wiedergegeben werden, wie Abbildung 2.3 zeigt. Im Gesprochenen wird zwischen gespannten und ungespannten Vokalen bzw. langen und kurzen Vokalen differenziert. Der Buchstabe ⟨a⟩ kann z. B. als [ɑ] (Langvokal wie in Pfad) oder [a] (Kurzvokal, wie in kalt) gesprochen werden (vgl. Abbildung 2.3). Die Schrift muss diese Unterscheidung mit einem reduzierten Grapheminventar ermöglichen, denn die Vokalquantität wird nicht über das Graphem selbst angezeigt; systematische Ausnahme bildet das ⟨ie⟩ (Eisenberg, Spitta & Voigt, 1994, S. 17; Eisenberg, 1989, S. 67).

Silbenstrukturelle Informationen, so Eisenberg, erlauben es dem Leser dennoch, die Vokalquantität zu erkennen. Durch den Silbenendrand und seine Beziehung zum Kern können sich Vokalgrapheme[25] auf Lang- und Kurzvokale beziehen, ohne dass dies beim Lesen zu Verwirrungen führt (Eisenberg, 1989, S. 69). Eisenberg formuliert dafür folgende Regel: Durch den Silbenschnitt werden Kurz- und Langvokale voneinander unterscheidbar (Eisenberg et al., 1994, S. 18). Der Vokal ist lang, wenn die betonte Silbe offen ist, also wenn auf den Vokal im Endrand der Silbe kein Konsonant folgt, wie z. B. bei Ro-se und le-sen oder bei schön (wegen schö-ne) und lebt (wegen le-ben). Er ist kurz, wenn die betonte Silbe geschlossen ist, also wenn sich mindestens ein Graphem im Endrand der Silbe befindet, wie z. B. bei Dan-ke und ler-nen oder bei Stift (wegen Stif-te) und Zelt (wegen Zel-te) (Eisen-

[25] Diphthonge sind davon ausgenommen, da hier keine Unterscheidung zwischen gespannt und ungespannt existiert (Eisenberg, 2006a, S. 312; Fuhrhop, 2006, S. 15).

| gespannte Vokale | ungespannte Vokale |
|---|---|
| [i] →⟨ie⟩ [ʃpiːs] – ⟨Spieß⟩ | [ɪ] →⟨i⟩ [ʃplɪnt] – ⟨Splint⟩ |
| [y] →⟨ü⟩ [tyːʀ] – ⟨Tür⟩ | [ʏ] →⟨ü⟩ [gəˈʀʏst] – ⟨Gerüst⟩ |
| [e] →⟨e⟩ [veːk] – ⟨Weg⟩ | [ɛ] →⟨e⟩ [vɛlt] – ⟨Welt⟩ |
| [ø] →⟨ö⟩ [ʃøːn] – ⟨schön⟩ | [œ] →⟨ö⟩ [ˈgœnən] – ⟨gönnen⟩ |
| [æ] →⟨ä⟩ [ˈtʀæːgə] – ⟨träge⟩ | [a] →⟨a⟩ [kalt] – ⟨kalt⟩ |
| [ɑ] →⟨a⟩ [pfɑːt] – ⟨Pfad⟩ | [ɔ] →⟨o⟩ [fʀɔst] – ⟨Frost⟩ |
| [o] →⟨o⟩ [ʃʀoːt] – ⟨Schrot⟩ | [ʊ] →⟨u⟩ [kʊnst] – ⟨Kunst⟩ |
| [u] →⟨u⟩ [huːt] – ⟨Hut⟩ | |
| Schwa | |
| [ə] →⟨e⟩ [ˈzɔnə] – ⟨Sonne⟩ | |

Abbildung 2.3: Vokalische Graphem-Phonem-Korrespondenzregeln (nach Eisenberg, 2006c, S. 69)

berg, 2006c, S. 72 f.).[26] In Einsilbern können die Regularitäten über die Langformbildung ermittelt werden (Eisenberg et al., 1994, S. 18). Im Folgenden werden die spezifischen Schärfungs- und Dehnungsgraphien und ihr Zusammenhang zur Vokalquantität erläutert.

Die wichtigste Schärfungsgraphie stellt nach Eisenberg (2006a, S. 313 f.) die Markierung von geschlossenen Silben durch die Verdopplung von Konsonantengraphemen (z. B. ⟨ll⟩ oder ⟨nn⟩ in Zelle oder Sonne) dar, die einen regelhaften Sonderfall der geschlossenen Silbe bildet. Es handelt sich hierbei um Silbengelenke (ambisilbische Konsonanten), da sie zwei Silben miteinander verbinden (Eisenberg et al., 1994, S. 18; Eisenberg, 1989, S. 78 ff.). Sie treten zwischen einer betonten Silbe mit ungespanntem Vokal und einer darauffolgenden unbetonten Silbe mit ungespanntem Vokal auf (Eisenberg, 2011, S. 90; Eisenberg, 2006a, S. 314). Sie sind Konsonanten, die zu beiden Silben gehören, und zwischen denen die Silbengrenze liegt (Butt & Eisenberg, 1990, S. 56 f.). Phonologisch entspricht damit ein Konsonant einer graphematischen Konsonantengemination. Die Verdopplung ist laut Eisenberg nicht an den Kurzvokal gebunden, sondern an bestimmte Kurzvokale, auf die Konsonanten folgen, die verdoppelt werden, weil sie Silbengelenke darstellen (Eisenberg, 2006a, S. 313; Eisenberg et al., 1994, S. 18). Nicht alle Silbengelenke sind durch eine Geminate gekennzeichnet. Bei den Bi- und Trigraphemen ⟨ng⟩, ⟨tz⟩, ⟨ck⟩, ⟨ch⟩ und ⟨sch⟩ findet keine Verdopplung statt (Eisenberg, 2006a, S. 314; Eisenberg, 2006c, S. 77 f.; Fuhrhop, 2006, S. 20).

Unter den Bereich der Dehnungsgraphien fällt das Dehnungs-h, welches keinem Phonem entspricht. Es ist ein stummes ⟨h⟩ – also ein Graphem, das kein Äquivalent im Lautlichen besitzt – weshalb Eisenberg phonologische Erklärungen der Schreibung negiert: „Man kann noch so genau artikulieren, ein Unterschied zwischen ⟨Boote⟩ und ⟨Bote⟩ ist nicht zu hören, ebenso wenig wie zwischen ⟨mahlen⟩ und ⟨malen⟩.“ (Eisenberg et al., 1994, S. 18) Das

[26] Es gibt auch wenige Ausnahmen, die nicht diesen Silbenregeln entsprechen (Keks, Mond). Sie gehören dem Peripheriebereich an und werden u. a. in Fuhrhop (2006, S. 14 ff.) erläutert.

Dehnungs-h kann dann gesetzt werden, wenn einem Vokalgraphem ein Sonorantengraphem folgt, also vor ⟨r⟩, ⟨l⟩, ⟨n⟩ und ⟨m⟩ (Eisenberg, 2006c, S. 73). Die Regel besagt nur, wo es stehen kann, wie z. B. in den Wörtern Lehrer, zählen, dehnen, aber nicht, wo es stehen muss, wie z. B. bei Blumen (Eisenberg, 1989, S. 74 f.). Das Dehnungs-h tritt in ungefähr jedem zweiten der möglichen Fälle auf (Eisenberg, 2011, S. 89; Eisenberg, 2006a, S. 317). Da Sonoranten häufig komplexe Endränder einleiten, markieren sie ungespannte Silbenkerne, wie z. B. bei Welt oder Furcht (vgl. hierzu das allgemeine Silbenbaugesetz und die Sonoritätshierarchie in Eisenberg (2006a, S. 103 f.) sowie die Schwereskala in Butt und Eisenberg (1990, S. 44)). Über das Dehnungs-h wird speziell hervorgehoben, dass der Kern trotz Sonorant gespannt ist. Laut Eisenberg ist die Setzung eines Dehnungs-h eigentlich überflüssig, da die Vokale in den Wörtern, in denen es vorkommt, ohnehin lang gesprochen werden würden (Eisenberg, 1989, S. 75). Eisenberg sieht im Dehnungs-h ebenfalls den Sinn in der Unterstützung beim Leseprozess (Eisenberg et al., 1994, S. 18; Butt & Eisenberg, 1990, S. 55 f.).

In der Graphematik wird das Dehnungs-h vom silbeninitialen h unterschieden. Es ist ebenfalls ein stummes ⟨h⟩, welches nicht zwingend notwendig ist (Eisenberg, 2006a, S. 315 f.). Im Gegensatz zum Dehnungs-h lassen sich für das silbeninitiale h nicht notwendige, sondern hinreichende Bedingungen formulieren, in denen es auftritt, weshalb es dem Kernwortschatz der Schreibung angehört (Eisenberg, 2011, S. 90; Eisenberg, 2006a, S. 316). Es wird zwischen zwei aufeinanderfolgenden Vokalen in einer betonten offenen Silbe und vor einer unbetonten (Schwa-)Silbe gesetzt (wie z. B. sehen, Schuhe, früher, nähen, ziehen) (Fuhrhop, 2006, S. 21 ff.).[27] Die Silbengrenze ist im Geschriebenen für den Leser durch das ⟨h⟩ segmental deutlicher markiert als im Gesprochenen (Butt & Eisenberg, 1990, S. 56). Zusätzlich wird durch das silbeninitiale h die Häufung von Vokalbuchstaben vermieden (Eisenberg, 2006a, S. 315). Die Schreibung *geen wird zugunsten von gehen verhindert: „Bemerkenswert an diesem Zug des Schriftsystems ist, dass es funktional wohl nützlich, aber nicht zwingend ist und dennoch mit hoher Regelmäßigkeit realisiert wird. Es entstehen regularisierte Muster, die die visuelle Analyse erleichtern.“ (Eisenberg, 2006a, S. 316)

Doppelte Vokalbuchstaben (⟨aa⟩, ⟨ee⟩, ⟨oo⟩) bilden die dritte große Gruppe an Dehnungsgraphien (Eisenberg, 2006a, S. 317). Sie werden in offenen Silben (Klee, Tee) oder vor ⟨r⟩, ⟨l⟩ und ⟨t⟩ (Meer, Saal, Boot) gesetzt. Analog zum Dehnungs- und silbeninitialen h ist die Vokalgemination als Längenanzeiger, so Eisenberg, redundant. Sie tritt ausschließlich an den Stellen auf, wo eine Schreibsilbe ohnehin lang zu lesen ist (Eisenberg, 2011, S. 89; Eisenberg, 1989, S. 76). Sie dienen als visuelle Markierung und werden vor Graphemen gesetzt, die sich üblicherweise ungespannten Vokalen anschließen, da sie zumeist komplexe Endränder einleiten. Die Markierung des Langvokals dient also auch hier dem Leser als eine Art Hilfe (Eisenberg, 1989, S. 75 f.).

[27] Das silbeninitiale h steht nicht bei Diphthongen (freuen); Ausnahme stellen mehrsilbige Grundformen mit ⟨ei⟩ dar (Weihe, leihen).

*Morphologisches Prinzip* Die Morphemunveränderlichkeit oder -konstanz im Geschriebenen tritt auf vielfältige Weise in Erscheinung, u. a. bei (Eisenberg, 2006c, S. 79 ff.):

- Tilgung von Lauten an der Morphemgrenze (Fah⟨rr⟩ad)
- Umlautkonstanz (R⟨ä⟩der)
- Verdopplung von Vokalgraphemen (gel⟨ee⟩rt)
- Dehnungs-h (gezä⟨h⟩lt)
- silbeninitialem h (sie⟨h⟩t)
- Gelenkschreibung (er re⟨nn⟩t)
- Auslautverhärtung (Ra⟨d⟩)

Das morphologische Prinzip überlagert das phonographische Prinzip. Diese Abweichungen bezeichnet Eisenberg (2006c, S. 79) als geregelt und funktional. Im Gesprochenen verändern Morpheme bei Flexion und Ableitung zuweilen ihre Lautgestalt, während eine Wortform im Geschriebenen relativ unverändert bleibt. Silbenschriftliche Informationen werden dabei vererbt, indem beispielsweise das silbeninitiale h in sieht durch sehen erhalten bleibt, oder die Konsonantengemination ⟨nn⟩ in rennt wegen rennen (Eisenberg, 2011, S. 93 f.; Fuhrhop, 2006, S. 27 ff.). Somit kommen immer wiederkehrende Zusammensetzungen von Graphemen in Wortformen vor, weshalb ein Leser Morpheme schnell wiederkennen und verarbeiten kann (Eisenberg, 2011, S. 92; Eisenberg, 2006c, S. 79). Durch die Analyse und Ableitung des prototypischen Zweisilbers (trochäischer Zweisilber: Abfolge von betonter und unbetonter Silbe) können grundlegende wortspezifische Informationen entdeckt und Zusammenhänge innerhalb des Schriftsystems ersichtlich werden (Eisenberg, 2011, S. 93). „Wörter mit ausschließlich einsilbigen Formen gibt es in den großen Klassen des Kernwortschatzes nicht, und den zweisilbigen Formen lässt sich ganz allgemein vieles über die Schreibung der einsilbigen Form entnehmen, z. B. ⟨lag⟩ mit ⟨g⟩ wegen ⟨lagen⟩, ⟨grob⟩ mit ⟨b⟩ wegen ⟨grobes⟩, ⟨Hund⟩ mit ⟨d⟩ wegen ⟨Hunde⟩.“ (Eisenberg et al., 1994, S. 18)

*Wortbildung und Wortformenbildung als Teil der Morphologie* Die Morphologie beschäftigt sich ebenfalls mit dem Aufbau von Wörtern und dem Prozess der Bildung neuer Wörter (Eisenberg, 2011, S. 94; Eisenberg, 2006a, S. 150, 209). Dabei ist die Wortbildung von der Wortformenbildung zu differenzieren (Fuhrhop, 2006, S. 33). Hauptfunktion der Wortbildung ist die Wortschatz- und Lexikonerweiterung (Eisenberg, 2006a, S. 211). Eisenberg (2006a, S. 209) unterscheidet die üblichen vier Regularitäten des Baus von Wörtern: Komposition, Präfigierung, Suffigierung und Konversion. Bei der Komposition treffen zwei Stammformen aufeinander (Kaffee+bohnen), bei der Präfigierung wird dem Stamm ein Wortbildungsaffix vorweggestellt (er+leben) und bei der Suffigierung angefügt (herz+lich). Keine Wörter oder Wortbilungselemente werden hingegen bei der Konversion hinzugefügt; der Stamm bleibt unverändert und wird in eine andere Wortart umgesetzt (Verb und Substantiv: schreiben – das Schreiben, Adjektiv und Substantiv: nett – der Nette) (Eisenberg, 2006a, S. 209 f.).

Bei der Wortformenbildung werden Deklination, Konjugation und Komparation unterschieden. Die Deklination bezieht sich auf die Flexion von Substantiven, Adjektiven, Pronomina und Artikeln, die Konjugation auf die von Verben. Komparation bezeichnet die Steigerung von Adjektiven, wobei es drei Steigerungsstufen gibt (Positiv: schnell, Komparativ: schneller und Superlativ: schnellste). Die Flexion ist zentraler Bestandteil der Grammatik des Deutschen, bei dem der Stamm eines Wortes um ein Flexionsmorphem ergänzt wird und eine grammatische Kategorie (wie Kasus, Numerus und Genus bei Substantiven) anzeigt, z. B. die/die/der Autos (Nominativ/Akkusativ/Genitiv, Plural, Neutrum) oder (sie ist Besitzerin) des Autos (Genitiv, Singular, Neutrum) (Eisenberg, 2006a, S. 150 f.).

*Wortübergreifendes Prinzip* Die Funktion der Großschreibung sieht Eisenberg in der Gliederung von Sätzen und der Orientierung in Texten. Sie dient der schnellen Informationsentnahme durch das Auge und erleichtert damit das Lesen (Eisenberg, 2006a, S. 344; Eisenberg & Feilke, 2001, S. 9). Die satzinterne Groß- und Kleinschreibung[28] wird bei Eisenberg nicht als isolierter Bestandteil gesehen, sondern im Kontext des Satzes. Die Entscheidung, ob ein Wort groß- oder kleingeschrieben wird, richtet sich nach der Funktion, die es innerhalb eines Satzes einnimmt (Eisenberg, 2006a, S. 348 ff.; Eisenberg et al., 1994, S. 19). So können beispielsweise tanzen bzw. Tanzen oder schöne bzw. Schöne je nach Kontext, in dem die Wörter eingesetzt werden, groß oder klein geschrieben werden.

Bei der syntaktischen Bestimmung von Substantiven oder substantivierten Wörtern werden Nominalgruppen in Sätzen bestimmt, die einen Wortverband darstellen, und immer einen Kern aufweisen. Den Kern stellt das Substantiv dar, das groß geschrieben wird und „Bedeutungszentrum“ ist (Eisenberg, 2006a, S. 345; Eisenberg & Feilke, 2001, S. 9). Die Kerne können als Subjekte (das Buch ist neu), Objekte (sie liest das Buch), Bestandteile von Präpositionalgruppen (sie liest in dem Buch), Genitivattribute (das Buch meiner Schwester; das Attribut nach dem Kern bildet selber eine Nominalgruppe mit Kern) und Adverbiale (sie liest den ganzen Tag) syntaktische Funktionen haben (Fuhrhop, 2006, S. 48; Maas, 1989, S. 164 f., S. 205 ff.).

Wörter aller Wortarten können großgeschrieben werden, wenn sie den Kern einer Nominalgruppe abbilden. Dies zeigt folgender Beispielsatz von Günther und Nünke (2005, S. 10), bei dem von der Wortart her kein Substantiv vorhanden ist, aber ein Verb, ein Adjektiv und ein Pronomen großgeschrieben werden: „Zum Tanzen trug sie das kleine Schwarze nicht so gerne, weil es zu ihrem neuen Ich nicht zu passen schien.“ Als Kern sind sie das semantische Zentrum einer Gesamtkonstruktion, die durch vorgestellte attributive Expansion erweitert werden können (Eisenberg, 2006a, S. 345). Die Attribuierbarkeit des Kerns zeichnet ihn als solchen aus und kann über Erweiterungen identifiziert werden (Günther & Nünke, 2005, S. 10; Röber-Siekmeyer, 1999, S. 60 ff.; Maas, 1989, S. 161 ff.). Diese Erweiterungsregel bestätigt die Großschreibung in dem vorherigen Beispielsatz: Zum *feierlichen* Tanzen trug sie das kleine *auffällige* Schwarze nicht so gerne, weil es zu ihrem neuen *bedachten* Ich nicht zu passen schien.

[28] Unberücksichtigt bleiben hier Erläuterungen zur Interpunktion, da sie keine Untersuchungseinheit bei der Auswertung des SRT in ZuRecht darstellt.

### Hinneys graphematisch basierte Rechtschreibdidaktik

Eine Umsetzung der Graphematik nach Eisenberg für die Rechtschreibdidaktik leistete Hinney (1997). Sie formulierte Lerninhalte und -ziele in Anlehnung an das phonographische, silbische und morphologische Prinzip nach Eisenberg. Diese sollten, so Hinney (2004, S. 76), zunächst im Fokus bei Schriftlernenden stehen, bevor der Peripheriebereich thematisiert wird. Für die Rechtschreibaneignung ist nach Auffassung von Hinney und Menzel (1998, S. 264) der lautorientierte Ansatz nicht aus- und hinreichend. Um Wortschreibungen herzuleiten, eignet sich dieser nicht als alleiniges Mittel zur richtigen Schreibung, da viele Abweichungen in Form von Sonderregelungen und Ausnahmeschreibungen formuliert werden müssen (Hinney, 2011, S. 205; Hinney & Menzel, 1998, S. 270). Zudem richten sich die Regeln nach dem Phonem-Graphem-Korrespondenz-Prinzip in ihrer Erklärung orthografischer Zusammenhänge an Schriftkundige (Hinney, 2011, S. 206). Hinney und Menzel (1998, S. 272 ff.) weisen darauf hin, dass eine große Schwierigkeit beim Aufbau von Rechtschreibfähigkeiten bei Kindern darin besteht, dass sie die Lang- und Kurzvokale nicht hören, was jedoch bei den von der Mündlichkeit ausgehenden didaktischen Modellen die Grundlage des Lernens bildet. Es wird also ein Wissen vorausgesetzt, über das Hinneys Ansicht nach Schriftlernende nicht verfügen (Hinney, 1997, S. 80). „Der Einsatz der bloßen Mündlichkeit zur Lösung rechtschreiblicher Probleme ist theoretisch und praktisch einfach unbrauchbar. Die Annahme, lautgetreue Wortschreibungen würden ohne eine ungefähre Einsicht in den Aufbau von Wörtern überhaupt möglich sein, kann nur von dem vertreten werden, der die Wortschreibung kennt.“ (Hinney & Menzel, 1998, S. 272)

Hinney folgt dem Ansatz Eisenbergs, Wortschreibungen über das phonographische, silbische und morphologische Prinzip zu beschreiben. Insbesondere stellt sie dabei den Stellenwert des silbischen Prinzips für eine einsichtige Darstellung der phonologischen Wortschreibungen als sogenannten „missing link“ heraus (Hinney, 1997, S. 89, 95 f.). Als Strategie für die Analyse von rechtschreiblichen Zusammenhängen entwickelt Hinney (1997, S. 131 ff.) das *zweischrittige Konstruktionsprinzip* der Wortschreibung vor dem Hintergrund des Eisenbergschen Orthografiemodells. Mit dem zweischrittigen Konstruktionsprinzip sollen Kindern sprachanalytische Untersuchungsmethoden an die Hand gegeben werden, um den Aufbau von Schreibungen selbst entdecken zu können (Hinney, 2004, S. 74). Basis für das Prinzip bilden die Regularitäten der Schreibsilbe im prototypischen Zweisilber (zweisilbige Langform) nach Eisenberg. Über ihn, so Hinney, können fast alle Merkmale der Wortschreibung (z. B. Vokallänge und -kürze, Silbengelenk, silbeninitiales h, ⟨ie⟩-Schreibung) erklärt werden, wie Tabelle 2.7 zeigt (Hinney, 2011, S. 214; Hinney, 2004, S. 80; Hinney & Menzel, 1998, S. 292).

Im ersten Schritt des zweischrittigen Konstruktionsprinzips sollen die phonologischen Gesetzmäßigkeiten (phonographisches und silbisches Schreiben) aufgedeckt werden (Hinney, 1997, S. 131 ff.). Dies soll über eine Silbenprobe des prototypischen Zweisilber erfolgen (Hinney, 2004, S. 79; Hinney & Menzel, 1998, S. 291 f.). Auf diese Weise analysieren Kinder die geschriebene und gesprochene Form und können z. B. den Unterschied in der Vokallänge bzw. -kürze der Phoneme /ʊ/ und /uː/ in Mutter und Musik durch den Silben-

| | unmarkiert | markiert |
|---|---|---|
| **Geschlossene Silbe** | brem-sen, Wel-ten, stol-ze | bren-nen, Mut-ter, hel-le |
| **Offene Silbe** | be-ten, Ho-se, gro-ße | lie-ben, Rie-se, Lie-be, Möh-ren, Koh-le, Haa-re, se-hen, Flö-he |

Tabelle 2.7: Prototypische Wortschreibungen und rechtschriftliche Besonderheiten (nach Hinney, 2004, S. 76)

schnitt bei Mut-ter und Mu-sik herleiten. Im zweiten Schritt sollen die Gesetzmäßigkeiten des Wortaufbaus und des morphologischen Prinzips aufgedeckt werden. Vererbte silbenstrukturelle Informationen werden so sichtbar (Hinney, 2004, S. 79 ff.; Hinney, 1997, S. 131 ff.). Auf diese Weise gelingt es Schriftlernenden herauszufinden, dass das flektierte Wort renn mit ⟨nn⟩ geschrieben wird, wegen ren-nen, oder der Einsilber Kind mit ⟨d⟩ wegen Kinder. Unterrichtspraktische Umsetzungen und qualitative Ergebnisse des Rechtschreiberwerbs mit dem zweischrittigen Konstruktionsprinzip der Wortschreibung sind u. a. in Hinney (2010, S. 82 ff.; 2004, S. 82 ff.; 1997, S. 137 ff.) dargestellt.

### Das Kompetenzmodell des SRT auf Basis graphematischer Forschung

Die Konzeption des SRT integriert die Prinzipien der Wortschreibung von Eisenberg und schließt sich dem didaktischen Ansatz von Hinney an. Eine Darstellung zur sprachsystematischen Rechtschreibdidaktik findet sich in Blatt (2010). Hier soll im Folgenden insbesondere die Modellierung von Rechtschreibkompetenz mit dem SRT beschrieben werden. Das Modell der Rechtschreibkompetenz, das Blatt und Voss vorlegen, weist unter Einbeziehung der Grundprinzipien der Wortschreibung und des wortübergreifenden Prinzips nach Eisenberg die folgenden fünf Teilkompetenzen aus (Blatt et al., 2011, S. 237):

- Phonographisch-silbisches Prinzip im Kernbereich
- Morphologisches Prinzip im Kernbereich
- Peripheriebereich
- Prinzip der Wortbildung
- Wortübergreifendes Prinzip

Das Rechtschreibmodell unterscheidet einen Kernbereich von einem Peripheriebereich. Wortschreibungen aus dem Kernbereich können nach Blatt und Voss über das phonographische, silbische, morphologische und wortübergreifende Prinzip regelhaft hergeleitet werden (Voss et al., 2007, S. 17). Eisenbergs phonographisches und silbisches Prinzip werden in dem Modell zur Teilkompetenz phonographisch-silbisches Prinzip zusammengefasst, worüber die phonologische Wortform erschlossen werden kann. Eisenbergs morphologisches Prinzip wird hingegen in die zwei Teilkompetenzen morphologisches Prinzip

| Orientierung an Prinzipien | Teilkompetenzen |
|---|---|
| Phonographisches und silbisches Prinzip im Kernbereich | Bezug herstellen zwischen Schrift- und Lautstruktur unter Berücksichtigung der silbenstrukturellen Informationen (Silbenanfangs- und -endrand und Silbenschnitt) |
| Morphologisches Prinzip im 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 |

Tabelle 2.8: Das Rechtschreibkompetenzmodell des SRT (nach Blatt, Voss, Kowalski & Jarsinski, 2011, S. 237)

im Kernbereich und Wortbildungsprinzip differenziert. Die Unterteilung basiert auf den unterschiedlichen Anforderungen bei der Verschriftung (Blatt et al., 2011, S. 238). Bei erstgenannter Teilkompetenz müssen über die Bildung des prototypischen Zweisilbers vererbte silbenschriftliche Informationen in Einsilbern und flektierten Wörtern aufgedeckt werden. Die zweitgenannte Teilkompetenz erfordert das Wissen und die Fähigkeit der Analyse von Stamm- und Wortbildungsmorphemen zur Schreibung abgeleiteter und zusammengesetzter Wörter (Blatt & Frahm, 2013, S. 15; Blatt et al., 2011, S. 238; Blatt, 2010, S. 109). Bei dem Prinzip der Wortbildung erfolgt keine Unterscheidung in einen Kern- und Peripheriebereich, da, so Blatt (2010, S. 109), die Wortbildung in beiden Bereichen gleich geregelt ist. Vom Kernbereich abgetrennt ist der Peripheriebereich. Er umfasst Wörter, deren Schreibung die Setzung von Dehnungs-h und Vokalgraphemverdopplung erfordern. Ebenfalls fallen Ausnahmeschreibungen, bei denen die Silbenregeln nicht greifen (z. B. Tiger, Papst), Fremdwörter (z. B. Browser, Smoothie, Rendezvous) sowie nicht ableitbare Einsilber (z. B. ab, wie, ihr) unter diese Teilkompetenz. Wörter aus diesen Bereichen müssen überwiegend durch Auswendiglernen und Üben eingeprägt werden (Blatt et al., 2011, S. 238; Blatt, 2010, S. 109; Voss et al., 2007, S. 17 f.). In Tabelle 2.8 sind die Teilkompetenzen[29] durch die unterschiedlichen Anforderungen der jeweiligen Prinzipien beschrieben. Sie werden von Blatt und Voss nicht im Sinne eines Stufenmodells verstanden. Stattdessen sollten sie parallel gefördert werden. Lediglich der Peripheriebereich nimmt eine Sonderstellung ein, da zunächst die Prinzipien des Kernbereichs gefestigt werden sollten, bevor auf Grundlage der dort gewonnenen Einsichten eine Transferleistung für die Ausnahmeschreibungen erbracht werden kann (Blatt, 2010, S. 113; Voss et al., 2007, S. 18).

[29] Sie werden im Kontext der Datenskalierung auch als Subskalen oder Kompetenzdimensionen bezeichnet.

### Wörter und Indikatoren des Tests

Der SRT wurde als zusammenhängender Text, als Wörtertest sowie kombinierter Wort-/Satztest konzipiert. Das Fließtextformat soll, so die Testautoren, kein Plädoyer für den Einsatz des Diktats im Unterricht sein (Blatt & Jarsinski, 2009, S. 96; Voss et al., 2007, S. 19). Es bietet aber im Rahmen von Lernstandserhebungen vielfältige rechtschriftliche Informationen (Blatt & Frahm, 2013, S. 19 f.; Blatt et al., 2011, S. 238). So können z. B. syntaxabhängige Schreibungen, wie die Groß-, Klein-, Getrennt- und Zusammenschreibung, erhoben, Wörter unterschiedlicher Wortart erfasst sowie gleiche Wörter in verschiedenen Kontexten und Wortformen analysiert werden (Blatt et al., 2011, S. 238 f.; Blatt, 2010, S. 118; Voss et al., 2007, S. 19). Als Nachteil wird im Vergleich zu anderen Testformaten der Zeitumfang im Hinblick auf die Testdurchführung und Dateneingabe ausgewiesen (Blatt & Frahm, 2013, S. 19; Blatt et al., 2011, S. 241). In der Ergänzungsstudie Orthografie zum *Hamburger Leseförderprojekt* (HeLp, s. u.) wurde daher ein Test eingesetzt, der sowohl das Schreiben einzelner Sätze als auch das Eintragen einzelner Wörter in vorgesehene Lücken beinhaltet (Frahm, 2012, S. 92 f.). Damit soll, so Blatt (2010, S. 118), eine zeitökonomische Studie gewährleistet werden, die gleichzeitig die Vorteile des Diktats besitzt. Dennoch zeigte sich durch den Einsatz in HeLp, dass die Teilkompetenz des wortübergreifenden Prinzips in dem Wörterdiktat nicht valide erhoben werden konnte, und in dem Wort/-Satztest keine vergleichbar hohe Reliabilität wie das Diktat aufwies (Blatt et al., 2011, S. 250).

Die Wortauswahl für den Einsatz in der Grundschule entstammt hauptsächlich dem Kernbereich (Blatt et al., 2011, S. 241). Daher wurden beispielsweise innerhalb des wortübergreifenden Prinzips fast ausschließlich Wörter ausgewählt, die aufgrund ihrer Wortart groß- bzw. kleingeschrieben werden. Der Test aus der IGLU-Voruntersuchung 2006 (112 Testwörter) weist mit dem Grundwortschatz von Bayern eine Überschneidung von 28 Wörtern auf (Voss et al., 2007, S. 20). Bezüglich des Tests, der in der IGLU-Haupterhebung 2006 und in ZuRecht eingesetzt wurde, schreibt Blatt, dass er insbesondere Wörter aus dem Grundwortschatz, aber auch ungebräuchliche Wörter beinhaltet, die mithilfe der in den Rahmenplänen vorgesehenen Regeln für die Grundschule zu erschließen sind (Blatt et al., 2011, S. 239; Blatt & Jarsinski, 2009, S. 96). Die Wörter, die über dieselben Prinzipien herleitbar sind, werden in einfacher, flektierter, abgeleiteter und zusammengesetzter Form in dem Diktat abgeprüft. Darüber soll getestet werden, ob die Schreibungen der Kinder regelgeleitet erfolgen und ein Wissenstransfer geleistet werden kann, oder von der Wortart oder dem Textzusammenhang beeinflusst sind (Blatt et al., 2011, S. 239).

Bei der Auswahl der Testwörter für die Sekundarstufe I des NEPS (s. u.) wurden die curricularen Anforderungen der jeweiligen Klassenstufe berücksichtigt (Blatt & Frahm, 2013, S. 29). In Jahrgangsstufe 5 lag der Fokus auf dem Kernbereich der Wortschreibung und die Großschreibung wurde im Rahmen des wortübergreifenden Prinzips erhoben. Ab Klasse 6 wurde die Anzahl der Struktureinheiten[30] des Peripheriebereichs ausgeweitet

[30] Hiermit sind die Analyseeinheiten gemeint, die im Folgenden auch als Indikatoren und im Rahmen der Testauswertung auch als Items bezeichnet werden.

und die Interpunktion miterfasst. Die dass/das-Schreibung als Teil des wortübergreifenden Prinzips wird ab der siebten Jahrgangsstufe berücksichtigt (Blatt & Frahm, 2013, S. 20 f.).

Innerhalb eines Wortes können mehrere Grapheme oder Graphemverbindungen Untersuchungseinheiten oder Struktureinheiten mehrerer Teilkompetenzen darstellen. Tabelle 2.9 zeigt einige beispielhafte Zuordnungen auf. So beinhaltet das Testwort schließlich Analyseeinheiten in den beiden Kernbereichen und in dem Prinzip der Wortbildung. Als Indikatoren dienen die Schreibung des komplexen Anfangsrandes (phonographisch-silbisches Prinzip), der regelhaft markierten Vokallänge ⟨ie⟩, von ⟨ß⟩ als stimmloses /s/ (wegen schlie-ßen, morphologisches Prinzip) sowie des Suffixes ⟨lich⟩ (Wortbildungsprinzip). Das Testwort Schnurrbarthaaren weist in allen fünf Teilkompetenzen Struktureinheiten auf: der komplexe Anfangsrand mit Kern ⟨Schnu⟩ (phonographisch-silbisches Prinzip), die Konsonantengraphemgemination bzw. das Silbengelenk ⟨rr⟩, das Stammmorphem ⟨bart⟩ (morphologisches Prinzip), die Vokalgraphemgemination ⟨aa⟩ (Peripherie), die Zusammen- (Prinzip der Wortbildung) sowie die Großschreibung (wortübergreifendes Prinzip). Bei Vergnügen wird der komplexe Anfangsrand, die offene unmarkierte Silbe und die Schreibung des stimmhaft gesprochenen Graphems ⟨g⟩ über die Struktureinheit ⟨gnüg⟩ (phonographisch-silbisches Prinzip), die Wortbildungsmorpheme bzw. Affixe ⟨ver⟩ und ⟨en⟩ (Wortbildungsprinzip) sowie die Großschreibung des Abstraktums (wortübergreifendes Prinzip) betrachtet. Bei einer Auswertung des phonographisch-silbischen und des Wortbildungsprinzips wäre die beispielhafte Schülerschreibung *vergnügen richtig, während diese im wortübergreifenden Prinzip als falsch bewertet werden würde. Wiederum wäre einem Kind bei der Schreibung *Ferknükn die Großschreibung gelungen und es würde ausschließlich Punkte innerhalb des wortübergreifenden Prinzips erhalten.

Da sich die Graphematik an den Schriftlernenden richtet und ihre Vertreter davon ausgehen, dass die Orthografie ein regelhaftes und erlernbares System darstellt, soll der Test Einsichten der Schriftlernenden in das Geschriebene erfassen (Blatt et al., 2011, S. 236). Daher werden die Wortschreibungen der Kinder „fähigkeitsbasiert“ ausgewertet und die Anzahl an richtig geschriebenen Struktureinheiten quantifiziert (Blatt et al., 2011, S. 243).

### 2.3.2 Studienergebnisse

Analyseergebnisse des SRT liegen bisher aus der IGLU-E Voruntersuchung 2006, der Ergänzungsstudie Orthografie zu HeLp sowie Entwicklungsstudien und Mode-Effekt-Studien innerhalb des NEPS vor. Dabei handelt es sich im Rahmen des IGLU-Pretests um Analysen in Klasse 4 sowie im Rahmen von HeLp und NEPS um Analysen in den Klassen 5 bis 7. Die Ergebnisse aus der Voruntersuchung zu IGLU umfassen eine Beschreibung des Kompetenzstandes der Schülerinnen und Schüler sowie eine Modellprüfung des SRT. Sie werden im Folgenden dargestellt sowie in den Abschnitten 4.1 und 4.2 in die Befunde aus ZuRecht eingeordnet.

In HeLp und NEPS erfolgten ebenfalls differenzierte Beschreibungen der Kompetenzen der Kinder und Überprüfungen des Kompetenzmodells. Da der SRT in den beiden Studien allerdings ein anderes Testformat besitzt, kaum bis keine Überschneidungen von Testwör-

| **Phonographisch-silbisches Prinzip im Kernbereich** | **Morphologisches Prinzip im Kernbereich** | **Peripheriebereich** | **Prinzip der Wortbildung** | **Wortübergreifendes Prinzip** |
|---|---|---|---|---|
| **schl**ießlich<br>#schl | schl**ieß**lich<br>#ie,<br>#ß | | schließ**lich**<br>#lich | |
| **gl**änzen<br>#gl,<br>#nzen | gl**ä**nzen<br>#ä | | | |
| **Schw**ä**n**e<br>#schw,<br>#n (falsch, wenn h oder nn oder äa, aa) | Schw**ä**ne<br>#ä | | | **S**chwäne<br>#S |
| ver**renken**<br>#renken | | | **verr**enken<br>#ver,<br>#rr | |
| Ver**gnüg**en<br>#gnüg | | | **Ver**gnüg**en**<br>#ver,<br>#en (falsch nd, ung) | **V**ergnügen<br>#V |
| | herrlich**e**<br>#e (Flexionsmorphem) | h**err**liche<br>#rr,<br>#e | herr**lich**e<br>#lich | |
| **fr**öhlich<br>#fr | fr**öh**lich<br>#ö,<br>#h | | fröh**lich**<br>#lich | |
| **ni**ckt<br>#ni | ni**ckt**<br>#ick (falsch ieck),<br>#t | | | |
| **Schnu**rrbarthaaren<br>#schnu | **Schnurrbart**haaren<br>#schnurr,<br>#bart | Schnurrbart**haar**en<br>#haar | Schnurrbartaaren<br>#Kompositum | **S**chnurrbarthaaren<br>#S |
| | | **Fohl**en<br>#f (falsch, wenn fpf, fph),<br>#ohl | | **F**ohlen<br>#F |
| qu**aken**<br>#aken | | **qu**aken<br>#qu | | |

Tabelle 2.9: Zuordnung ausgewählter Struktureinheiten zu den SRT-Teilkompetenzen (nach Blatt, Voss, Kowalski & Jarsinski, 2011, S. 240)

tern mit dem Test in ZuRecht aufweist und für die Sekundarstufe I konzipiert worden ist, werden diese Ergebnisse nicht aufgeführt. Hierbei soll auf die Veröffentlichungen von Blatt et al. (2011) für HeLp sowie Jarsinski (2014), Blatt und Frahm (2013) und Frahm (2012) für NEPS verwiesen werden. Eine Erprobung der sprachsystematischen Rechtschreibdidaktik fand mit Kindern einer Klasse in Jahrgangsstufe 2 in Hamburg statt. Erste Ergebnisse des Unterrichtsprojekts sind in Hinney und Pagel (2007) veröffentlicht.

### IGLU-Voruntersuchung

Die Voruntersuchung zu IGLU-E fand im Sommer 2005 statt. Es wurden 486 Grundschulkinder aus 5 Schulen mit jeweils 2 Klassen am Ende der vierten Jahrgangsstufe mit dem

SRT getestet. An der Erhebung beteiligten sich die fünf Bundesländer Baden-Württemberg, Berlin, Niedersachsen, Saarland und Sachsen (Kowalski, 2007, S. 62; Voss et al., 2007, S. 20). Der Test wurde im Fließtextformat eingesetzt und besteht aus 112 Wörtern und 103 Analyseeinheiten, die nach dem SRT-Kompetenzmodell ausgewertet wurden. Bei dem Inhalt des Diktates handelt es sich um eine kurze fiktive Geschichte, in der Tiere in der Natur miteinander kommunizieren (Voss et al., 2007, S. 18 f.).

Die Datenauswertung in der Voruntersuchung unterschied sich hinsichtlich der Transkribierung sowie Kodierung mit Hilfe des Programms „MaxQda“[31] von der in ZuRecht genutzten Vorgehensweise (vgl. Abschnitt 3.2). Die Dateneingabe erfolgte in der Voruntersuchung durch das Übertragen der Schülertests in Textdateien. Alle geschriebenen Wörter und Satzzeichen der Kinder und alle darin enthaltenen Fehler wurden abgetippt. In MaxQda wurden die Textdateien mit den transkribierten Diktaten eingelesen und weiterverarbeitet. Hier erfolgte eine computergestützte Inhaltsanalyse auf Gesamtwortebene und auf Basis der fünf Teilkompetenzen. Dafür wurden Diktionäre erstellt und die einzelnen Schreibungen auf Basis der Struktureinheiten ausgewertet, indem sie als richtig oder falsch markiert worden sind (Blatt et al., 2011, S. 243; Kowalski, 2007, S. 55 f.).

Die 112 Wörter des Diktats wurden in 2.442 Varianten geschrieben. Zwei Kinder haben den Text fehlerfrei verschriftet. Das schlechteste Ergebnis liegt bei 70 Falschschreibungen. Im Durchschnitt wurden 86,4 Wörter (84 Prozent) richtig geschrieben. Die differenzierte Auswertung der Teilkompetenzen ergab, dass 90 Prozent der Struktureinheiten des phonographisch-silbischen Prinzips richtig geschrieben wurden ($M = 52,9$; $SD = 5,6$). Es ist das Prinzip, das den Schülerinnen und Schülern am leichtesten fällt. Darauf folgen das Prinzip der Wortbildung und das wortübergreifende Prinzip mit jeweils 84 Prozent Richtigschreibungen ($M = 15$; $SD = 2,5$ bzw. $M = 22$; $SD = 3,9$). Im Durchschnitt werden 82 Prozent der Analyseeinheiten des morphologischen Prinzips richtig verschriftet ($M = 37,6$; $SD = 6,1$). Die meisten Probleme haben die Kinder mit dem Peripheriebereich. Hier werden 78 Prozent der Struktureinheiten richtig geschrieben ($M = 17,3$; $SD = 3,9$) (Voss et al., 2007, S. 22 f.).

Ebenfalls wurde die Struktur des Kompetenzmodells empirisch überprüft.[32] Die latenten Korrelationen sind in Tabelle 2.10 dargestellt. Der höchste Zusammenhang zeigt sich zwischen den beiden Kernbereichen (0,99). Der Peripheriebereich weist die jeweils niedrigsten Zusammenhänge mit den weiteren vier Teilkompetenzen auf. Die Korrelation mit dem phonographisch-silbischen Prinzip beträgt 0,76, mit dem morphologischen Prinzip 0,79, mit dem Wortbildungsprinzip 0,78 und mit dem wortübergreifenden Prinzip fällt sie mit einem Wert von 0,70 am kleinsten aus. Ein Vergleich dieses 5D-Modells mit einem eindimensionalen weist das theoriekonforme fünfdimensionale als vorzuziehendes Modell aus. Die Reliabilität aller Teilkompetenzen pendelt um 0,9, ist am niedrigsten für den

[31] MaxQda ist ein Programm zur computerunterstützten Analyse qualitativer und quantitativer Daten. Es ermöglicht systematische Analysen durch Kodierung von Textmengen (z. B. Interviewaufzeichnungen) und wird bevorzugt bei der qualitativen Inhaltsanalyse genutzt. Das Erweiterungsmodul „MaxDictio“ bietet MaxQda auch für eine quantitative Inhaltsanalyse an.

[32] Details zu den statistischen Berechnungen und zur Einordnung der Werte finden sich in Abschnitt 4.2.2. Dort werden die hier berichteten Ergebnisse ebenfalls mit den Werten aus ZuRecht verglichen.

| Teilkompetenzen | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| (1) Phonographisch-silbisches Prinzip im Kernbereich | | | | |
| (2) Morphologisches Prinzip im Kernbereich | 0,99 | | | |
| (3) Peripheriebereich | 0,76 | 0,79 | | |
| (4) Prinzip der Wortbildung | 0,97 | 0,96 | 0,78 | |
| (5) Wortübergreifendes Prinzip | 0,84 | 0,84 | 0,70 | 0,91 |

Tabelle 2.10: Latente Interkorrelationen des SRT in der IGLU-Voruntersuchung (in Anlehnung an Voss, Blatt & Kowalski, 2007, S. 24)

Peripheriebereich (0,87) und am höchsten für das Wortbildungsprinzip (0,97) (Voss et al., 2007, S. 23 ff.).

Eine qualitative Betrachtung ausgewählter Struktureinheiten der Teilkompetenzen (wie z. B. Silbengelenk- und Umlautschreibung, silbeninitiales h, Dehnungs-h oder Flexionsmorpheme) zeigte, dass es innerhalb der Kategorien auffällige Differenzen in der Aufgabenschwierigkeit gab. Beispielsweise fällt die Schreibung der Konsonantengemination in dem Wort ko⟨mm⟩en der Schülerschaft leichter (96 Prozent Richtigschreibungen) als in dem Wort Schnu⟨rr⟩-(barthaaren) (34 Prozent Richtigschreibungen). Die Umlautschreibung in dem Testwort n⟨ä⟩chsten wurde zu 84 Prozent richtig geschrieben, während sie in gl⟨ä⟩nzenden von rund der Hälfte der Kinder (56 Prozent) gelöst wurde. Die relative Lösungshäufigkeit des silbeninitialen h unterscheidet sich bei Re⟨h⟩ (95 Prozent Richtigschreibungen) und frö⟨h⟩lich (70 Prozent Richtigschreibungen) um 25 Prozentpunkte. Als Ursachen für diese Unterschiede im Grad der Aufgabenschwierigkeit wird zum einen der Bekanntheitsgrad von Wörtern, zum anderen die Komplexität der Wortstruktur bzw. die höhere Schwierigkeit bei flektierten, abgeleiteten und zusammengesetzten Wörtern genannt (Voss et al., 2007, S. 26 f.).

## 2.4 Theoretischer Vergleich der Tests aus ZuRecht

In den Abschnitten 2.2 und 2.3 wurden die beiden Rechtschreibtests gutschrift-diagnose und SRT jeweils in ihrem linguistisch basierten Aufbau und ihren Grundannahmen sowie ihrem Testdesign und -einsatz beschrieben. Darauf basierend sollen im Folgenden die wesentlichen schrifttheoretischen Unterschiede und Gemeinsamkeiten zwischen den beiden Tests in zwei Schritten aufgeschlüsselt werden. Zunächst werden in Abschnitt 2.4.1 über eine quantitative Inhaltsanalyse Überschneidungen in den Indikatorzuordnungen zu den Teilkompetenzen ausgezählt und analysiert. Eine vertiefende Analyse der Unterschiede und Gemeinsamkeiten der beiden Tests erfolgt in Abschnitt 2.4.2, indem die konzeptionellen Besonderheiten herausgestellt und interpretativ gegenübergestellt werden. Hierbei werden weitere einzelne Zuweisungen von Indikatoren zu den Teilkompetenzen

herausgegriffen und zum Teil kritisch diskutiert. Eine Auseinandersetzung mit den schriftsprachtheoretischen Begründungen der Auswahl und Zuweisung von Indikatoren ist bei Diagnoseinstrumenten erforderlich, da sie direkte Konsequenzen auf die Analyse von Schreibprodukten sowie auf Diagnostik und Fördermaßnahmen haben. Neben gutschrift und dem SRT werden zudem die HSP und AFRA (vgl. Abschnitt 2.1) wieder aufgegriffen. Im Rahmen eines Exkurses erfolgt in Abschnitt 2.4.3 eine Betrachtung ausgewählter Items der HSP und AFRA, die in Zusammenhang mit dem Peripheriebereich des SRT gesetzt werden. Für diese grundlegende und strukturell wichtige Teilkompetenz des SRT findet an dieser Stelle ein Vergleich der Analyseeinheiten statt.

### 2.4.1 Indikatorenzugehörigkeit

In den Tabellen 2.5 und 2.9 sowie den zugehörigen Absätzen für gutschrift-diagnose und den SRT wurden die Analyseeinheiten und deren Zuordnung zu den Teilkompetenzen erläutert. Nun soll die Indikatorenzugehörigkeit im Vergleich betrachtet werden. Dafür wurden Rechtschreibphänomene ausgewählt, die in beiden Tests Analyseeinheiten darstellen. Die Bezeichnungen der Phänomene basieren zumeist auf den verwendeten Begrifflichkeiten der Testkonzeptionen. Die Indikatoren besitzen dabei unterschiedliche Differenziertheitsgrade, da z. B. sowohl einzelne Grapheme (wie ⟨qu⟩, ⟨st⟩, ⟨sp⟩ oder Dehnungs-h) als auch Gruppen von Graphemen bzw. rechtschriftliche Merkmale (wie Affrikaten, Auslautverhärtung oder Flexionsmorpheme) aufgeführt sind. Sie ergeben sich aus den Analyseeinheiten, die von den Testautoren genannt werden. Zu berücksichtigen gilt weiterhin, dass die Teilkompetenzen unterschiedlich stark mit Indikatoren vertreten sind. Tabelle 2.11 zeigt das Ergebnis dieser Zusammentragung.

Größere Überschneidungen sind zwischen der elementaren phonographischen Teilkompetenz (gutschrift) und dem phonographisch-silbischen Prinzip (SRT) sowie zwischen der erweiterten grammatischen Teilkompetenz (gutschrift) und dem morphologischen Prinzip (SRT) erkennbar. Die elementare phonographische Teilkompetenz und das phonographisch-silbische Prinzip weisen bei der (allgemeinen) Aneignung von Phonem-Graphem-Korrespondenzen sowie bei der Schreibung von Vokalen, Diphthongen, Konsonanten – im Speziellen von Plosiven, Affrikaten und Nasalen – sowie von /ʃ/ vor /t/ und /p/ als ⟨st⟩ und ⟨sp⟩ gemeinsame Indikatoren auf. Bei dem Verschriften von ⟨st⟩ und ⟨sp⟩ am Morphembeginn werden von beiden Tests unterschiedliche Zugänge gewählt. Die Kinder sollen im SRT die Beziehung zwischen gesprochenen und geschriebenen Silbenanfangsrändern begreifen (Blatt, 2010, S. 104), während sie in gutschrift die verschiedene Verschriftung derselben Sprachlaute beherrschen sollen (Löffler & Meyer-Schepers, 2005, S. 83). Obwohl bei der Schreibung des /ʃ/ vor /t/ und /p/ die Position bedacht werden muss, also eine weitere kognitive Leistung zu erfolgen hat, indem sich die Schreibung nicht allein aufgrund der Lautform ergibt, wird dieser Bereich in gutschrift der elementaren phonographischen Teilkompetenz zugeordnet, was Bremerich-Vos (2004, S. 98) kritisch anmerkt. Unter die Affrikaten wird von gutschrift – entgegen gängiger Klassifizierungen (vgl. Abschnitt 2.2) – das Digraph ⟨qu⟩ gezählt, weshalb beide Rechtschreibeinheiten

in Tabelle 2.11 untereinander aufgeführt sind. Damit wird der Schreibung von ⟨qu⟩ eine ähnliche Schwierigkeit wie der von ⟨pf⟩ und ⟨z⟩ in z. B. den Testwörtern über⟨qu⟩eren, glän⟨z⟩en, em⟨pf⟩indlich zugewiesen (Valtin, Badel et al., 2003, S. 237). Im SRT erfolgt hingegen eine Zuordnung, aufgrund der aus den Ergebnissen der IGLU-Voruntersuchung formulierten Annahme, dass die Buchstabenverbindung den Kindern nicht geläufig ist, zu dem Peripheriebereich (Blatt et al., 2011, S. 240; Voss et al., 2007, S. 22). Zur Affrikate ⟨pf⟩ ist anzumerken, dass sie im SRT, je nach Funktion, bei zwei unterschiedlichen Teilkompetenzen betrachtet wird. In der Position als Silbengelenk stellt sie ein Item des phonographisch-silbischen Prinzips (z. B. hüpfen) dar und als Silbenanfangsrand ein Item des Peripheriebereichs (z. B. Pferd).

Gemeinsame Analyseeinheiten zwischen der erweiterten grammatischen Kompetenz und dem morphologischen Prinzip beziehen sich insbesondere auf die Verschriftung von orthografischen Phänomenen in flektierten und abgeleiteten Wörtern: die Schreibung des stimmlosen /s/ als ⟨s⟩ und ⟨ß⟩ nach langem Vokal, des silbenanlautenden h bzw. silbeninitialen h, des Dehnungs-e bzw. ⟨ie⟩ sowie von Doppelkonsonantengraphemen bzw. Silbengelenken. Bei der Analyseeinheit Auslautverhärtung findet sich in Tabelle 2.11 eine dreifache Zuordnung. Bei dem SRT wird die Auslautverhärtung innerhalb des morphologischen Prinzips, wie z. B. bei Pfer⟨d⟩ betrachtet. Sie wird in gutschrift-diagnose bei den beiden grammatischen Teilkompetenzen unter dem Aspekt der Ableitung der Verschlusslautung aufgeführt, da hier eine Unterscheidung zwischen elementaren und erweiterten Ableitungsoperationen stattfindet. So fällt Schrei⟨b⟩maschine in die elementare und empfin⟨d⟩lich in die erweiterte grammatische Teilkompetenz (Valtin et al., 2004a, S. 143 f.). Die unterschiedliche Verortung in die beiden Niveaus erscheint diskussionswürdig, da es sich bei beiden Testwörtern um – durch Komposition und Derivation entstandene – morphologisch komplexe Wörter handelt, bei denen jeweils die Auslautverhärtung durch den Prozess der Rückführung auf die Grundform (mit zusätzlichem Wechsel der Wortart) abgeleitet werden muss (Bremerich-Vos, 2004, S. 99).

Weniger bis keine Überschneidungen in den Analyseeinheiten sind bei den restlichen Teilkompetenzen vorhanden. Die Struktureinheiten des Peripheriebereichs verteilen sich über fast alle gutschrift-Subskalen. Auf diese breite Fächerung wird weiter unten im Rahmen des konzeptionellen Vergleichs und im Zusammenhang mit der Diskussion der Zuordnung von Analyseeinheiten zu den Teilkompetenzen eingegangen. Das wortübergreifende Prinzip und das Prinzip der Wortbildung weisen ausschließlich Überschneidungen mit den beiden grammatischen Teilkompetenzen auf. Im Bereich der Großschreibung erfolgt bei gutschrift eine Differenzierung in die Stufen elementar und erweitert, die im SRT nicht vorgenommen wird. Unter das Wortbildungsprinzip und die grammatischen gutschrift-Teilkompetenzen sammeln sich die Zusammen- und Getrenntschreibung sowie die Präfigierung und Suffigierung. Die Schreibung der Endungen ⟨e⟩, ⟨en⟩, ⟨el⟩ und ⟨er⟩ werden in beiden Tests separat ausgewiesen und unter der elementaren grammatischen Teilkompetenz bzw. dem phonographisch-silbischen Prinzip betrachtet. Hier zeigen sich gegensätzliche Annahmen, da eine grammatische Zugangsweise von gutschrift und eine laut- und silbenanalytische vom SRT gewählt werden. Unter Verwendung der *Explizitlautung* kann die zweisilbige Struktur und der zweite Silbenkern beispielsweise bei den

Wörtern geben und wollen erkannt werden.[33] Bei der Explizitlautung werden die Wortformen einzeln und so ausgesprochen, dass die artikulatorischen Merkmale jedes einzelnen Lautes sowie alle Silben und jeder Vokal als Silbenkern zu hören sind, wie z. B. [geːbən] und [vɔlən]. Die Explizitlautung ist von der *Überlautung* abzugrenzen, bei der „künstliche Lautgestalten“ entstehen. Hier wird u. a. Schwa durch einen Vollvokal ersetzt (vgl. die Aussprache zwischen [ʔɛntˈlaʊ̯fən] in Explizit- und [ʔɛntˈlaʊ̯fɛn] in Überlautung) (Eisenberg, 2006c, S. 51 ff.). Zu beachten ist weiterhin, dass sie bei Schriftlernenden nicht vorausgesetzt werden kann, sondern erst allmählich beherrscht wird (Hinney, 2010, S. 58).

### 2.4.2 Konzeptionelle Besonderheiten

#### Zugänge zu Wortschreibungen

Ein relevanter Unterschied in der Konzeption zwischen gutschrift-diagnose und dem SRT besteht in der Rolle der Silbe (Blatt, 2010, S. 123). Die Schreib- und Sprechsilbe ist in der theoretischen Konzeption, der testanalytischen Auswertung und der didaktischen Umsetzung des SRT entscheidend und fest verankert. In dem SRT-Kompetenzmodell wird Eisenbergs silbische Zugangsweise in der Teilkompetenz phonographisch-silbisches Prinzip abgebildet. Durch das phonographische Prinzip gepaart mit dem silbischen Prinzip lassen sich Schreibungen phonologisch begründen (Eisenberg, 2006a, S. 319). In dem zweischrittigen Konstruktionsprinzip nach Hinney bildet dies einen Grundgedanken zur Herleitung von Wortschreibungen (wobei der prototypische Zweisilber gebildet und die Silbengrenze bestimmt werden muss). In gutschrift sind die Schreibungen nach den phonographischen Kompetenzdimensionen in lautanalytischer Orientierung erschließbar und es werden innerhalb dieser Teilkompetenzen keine silbischen Zugangswege zur Schreibung von Wörtern dargestellt (Voss, Löffler et al., 2008, S. 144; Löffler & Meyer-Schepers, 2007, S. 184). Stattdesssen werden die Abweichungen von lauttreuen Schreibungen betrachtet und darauf basierend die beiden unterschiedlichen Fähigkeiten phonographisch und grammatisch definiert: „Jene Stellen, wo ein Rekurs auf die Lautebene zur Klärung der Schreibung nicht hilft, sind als Domänen der (wort-/satz-) grammatischen Kompetenz definiert.“ (Löffler & Meyer-Schepers, 2007, S. 184) Die Graphematik nach Eisenberg und der SRT verstehen sich als Vertreter der Interdependenzthese. Die Schrift- und Lautsprache werden hierbei als regelhaft aufeinander bezogene, aber selbstständige Systeme angesehen (Eisenberg, 1990, S. 6; Eisenberg, 1989, S. 58). Es wird damit von keiner direkten Ableitbarkeit der graphematischen Form aus der phonologischen ausgegangen (Hinney & Menzel, 1998, S. 270), sondern Wortschreibungen über das phonographische, silbische, morphologische und wortübergreifende Prinzip analysiert und rekonstruiert.

[33] Eisenberg (2006a, S. 130) schreibt, dass etwa 70 Prozent der Zweisilber des Deutschen als zweite eine Schwasilbe besitzen.

| | gutschrift | | | | SRT | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Teilkompetenzen/ Rechtschreibphänomene | Elementar phonographisch | Elementar grammatisch | Erweitert phonographisch | Erweitert grammatisch | Phonographisch-silbisches Prinzip im Kernbereich | Morphologisches Prinzip im Kernbereich | Peripheriebereich | Prinzip der Wortbildung | Wortübergreifendes Prinzip |
| Phonem-Graphem-Korrespondenzen | X | | | | X | | | | |
| Vokale, Diphthonge, Konsonanten | X | | | | X | | | | |
| Plosive | X | | | | X | | | | |
| Affrikaten | X | | | | X | | | | |
| Schreibung von <qu> | X | | | | | | X | | |
| Nasale | X | | | | X | | | | |
| Schreibung von /f/ als <f> und <v> | | X | | | X | | X | | |
| Schreibung von /ʃ/ vor /t/ und /p/ als <st> und <sp> | X | | | | X | | | | |
| Schreibung von /s/ als <s> oder <ß> | | | | X | X | | | | |
| Schreibung von /s/ als <s> oder <ß> in flektierten Wörtern | | | | X | | X | | | |
| silbenanlautendes h/ silbeninitiales h | | | | X | X | | | | |
| silbenanlautendes h/ silbeninitiales h in flektierten Wörtern | | | | X | | X | | | |
| Dehnungs-h | | | X | | | | X | | |
| Dehnungs-e bzw. <ie> | | | X | | X | | | | |
| Dehnungs-e bzw. <ie> in flektierten Wörtern | | | | X | | X | | | |
| Ableitung Umlaut | | X | | | | X | | | |
| Konsonantengraphemgemination bzw. Silbengelenke | | | X | | X | | | | |
| Konsonantengraphemgemination bzw. Silbengelenke in flektierten Wörtern | | | | X | | X | | | |
| Vokalgraphemgemination | | | X | | | | X | | |
| Auslautverhärtung | | X | | X | | X | | | |
| Flexionsmorpheme | | X | | | | X | | | |
| Zusammen- und Getrenntschreibung | | | | X | | | | X | |
| Präfigierung | | X | | | | | | X | |
| Suffigierung | | | | X | | | | X | |
| Endungen <e>, <en>, <el>, <er> | | X | | | X | | | | |
| Großschreibung Konkreta, Satzanfang | | X | | | | | | | X |
| Großschreibung Abstrakta, Nominalisierung | | | | X | | | | | X |
| Kleinschreibung | | X | | | | | | | X |

Tabelle 2.11: Vergleichende Zuordnung der Indikatoren zu den Teilkompetenzen

### Kürze- und Längenzeichen

Ferner werden viele rechtschriftliche Merkmale bei gutschrift anders verstanden und erklärt als innerhalb des silbenorientierten Konzepts des SRT. Der Konsonantengraphemverdopplung wird in gutschrift die Aufgabe zur Markierung der Kürze des Akzentvokals zugewiesen (Löffler & Meyer-Schepers, 2005, S. 92; Valtin, Badel et al., 2003, S. 239). In der Graphematik bzw. im SRT wird diese hingegen in der Kennzeichnung der syllabischen Struktur gesehen: „Steht in einer phonologischen Wortform zwischen einem betonten ungespannten und einem unbetonten Vokal ein einzelner Konsonant, so ist dieser Konsonant ein Silbengelenk. [...] Doppelkonsonantengrapheme haben nach dieser Auffassung ihren Ursprung nicht in der Kennzeichnung von Vokalkürze, sondern bei der Markierung von Silbengelenken." (Eisenberg, 2006c, S. 76 f.). In diesem Rahmen soll zusätzlich die entsprechende Erläuterung des amtlichen Regelwerks aufgeführt werden. Hier heißt es: „Folgt im Wortstamm auf einen betonten kurzen Vokal nur ein einzelner Konsonant, so kennzeichnet man die Kürze des Vokals durch Verdopplung des Konsonantenbuchstabens." (Rat für deutsche Rechtschreibung, 2006, S. 18). Die Ansichten und damit verbundenen Regelformulierungen zur Konsonantengraphemgemination und Vokalkürze unterscheiden sich, weisen aber doch gleiche Voraussetzungen zur Herleitung der korrekten Schreibung auf. Zum einen muss die Anzahl der Grapheme, die auf den Vokal folgen, bestimmt werden, wofür zum anderen strukturelles Wissen (Silbe bzw. Morphem) notwendig ist. So kann eine korrekte Entscheidung über die Schreibweise von z. B. den Wörtern kommt und Wand getroffen werden.

Bei dem Konzept von gutschrift dienen Dehnungs-h, Dehnungs-e, silbenanlautendes h und Doppelvokalgrapheme zur Markierung der langen Akzentvokale – parallel zur Funktion der Kürzezeichen zur Kennzeichnung kurzer Akzentvokale. Ihnen wird zudem die Aufgabe der Disambiguierung von homophonen Formen zugeschrieben (Valtin et al., 2004a, S. 142 f.; Valtin, Badel et al., 2003, S. 240). „Dehnungs- und Kürzezeichen sind im deutschen Sprachsystem phonologisch relevant, also Mittel der Wortunterscheidung. Insofern die Dauer in unbetonten Silben sprachlich ohne Funktion ist, ist die graphematische Repräsentation der Vokalquantität ausschließlich auf den Akzentvokal bezogen und der Verschriftungsort für Dehnungs- und Kürzezeichen auf die Position nach dem Akzentvokal festgelegt." (Löffler & Meyer-Schepers, 2005, S. 92) Die Anzahl homophoner Formen ist aber überschaubar und sie werden laut Eisenberg (2006c, S. 83 f.) als Mittel zur Differenzierung von geschriebenen Wörtern überschätzt, da u. a. eine Verwechslung der Wortformen durch den Kontext meist ausgeschlossen ist. G. Thomé (2000, S. 14 f.) vertritt die Ansicht, dass die Andersschreibung gleichlautender Wörter historisch bedingt ist und nur zufällig die Unterscheidung der Homonyme auf schriftlicher Ebene unterstützt.

Die Setzung von Dehnungs-h und doppelten Vokalbuchstaben sind nach Eisenberg (1989, S. 75 f.) als Längenanzeiger redundant und dienen dem Leser als eine Unterstützung und Erleichterung. Die leserfreundliche Funktion der Orthografie wird aus Sicht graphematischer Forschung immer wieder betont (Munske, 2005, S. 30 ff.; Eisenberg & Feilke, 2001, S. 8). Darüber erhält die Rechtschreibung nicht nur den Wert des normgerechten Schreibens,

sondern auch eine Bedeutung als Lerngegenstand. So schreiben Eisenberg et al. (1994, S. 15): „Zwar beginnen Kinder mit Schreibversuchen, bevor sie lesen können, aber das orthographische System ist auf eine möglichst große Erleichterung des Lesens ausgerichtet, nicht des Schreibens.“ Die graphematische Wiedergabe phonologisch kurzer und langer Vokale wird in gutschrift-diagnose also über die Setzung der entsprechenden Längen- und Kürzeanzeiger zur Markierung von Schärfungs- und Dehnungsgraphien erklärt. Bei dem SRT steht die Herleitung der Verschriftung kurz und lang gesprochener Vokale über den silbenstrukturellen Aufbau von Wortschreibungen im Fokus.

Auf die differenzierten Sichtweisen zwischen gutschrift-diagnose und dem SRT bei der Schreibung des Graphems ⟨h⟩ wird im Folgenden näher eingegangen. Es werden Dehnungs-h und silbenanlautendes h bzw. silbeninitiales h voneinander unterschieden. Dem Dehnungs-h wird in der gutschrift-Konzeption die Funktion zur Kennzeichnung langer Vokale zugeschrieben (Valtin, Badel et al., 2003, S. 239). Aus graphematischer Sicht wird die Setzung des Dehnungs-h, wie bereits ausgeführt, im Nutzen für das Lesen hervorgehoben (Eisenberg et al., 1994, S. 18). Blatt (2006) beschreibt gängige Ausschlussverfahren für die Schreibung von Wörtern mit Dehnungs-h. Diese Schreibhinweise geben allerdings nur an, wo das ⟨h⟩ stehen kann (vor ⟨l⟩, ⟨m⟩, ⟨n⟩ oder ⟨r⟩), aber nicht, wo es auftreten muss. Im SRT gilt das Dehnungs-h daher als nicht regelhaft herleitbar und wird dem Peripheriebereich zugeordnet (Blatt, 2010, S. 107). Dem Kernbereich wird hingegen das silbeninitiale h zugerechnet. Es wird Blatt (2010, S. 115) zufolge regelhaft gesetzt, wenn zwei Silbenkerne im Silbenschnitt aufeinander stoßen. Das silbeninitiale h wird als ein stummes h sowie silbentrennendes Zeichen, und nicht als ein Dehnungszeichen aufgefasst (Blatt, 2006; Eisenberg, 2006a, S. 315).

In gutschrift-diagnose dient das Dehnungs-h als Indikator in der erweiterten phonographischen Kompetenz und es wird nicht von Ausnahmeschreibungen im Kontext der h-Schreibung gesprochen. Diese Zuordnung zu einem phonographischen Bereich hinterfragt Bremerich-Vos (2004, S. 99), da die Schreibung mehr als eine Lautanalyse voraussetzt (vgl. die oben genannten notwendigen, aber nicht hinreichenden Bedingungen). Das silbenanlautende h stellt dagegen für die erweiterte grammatische Kompetenz in gutschrift eine Analyseeinheit dar. Es wird beschrieben als „‚stummen‘ -h, das erst silbenanlautend zum ‚Laut-h‘ wird“ (Löffler & Meyer-Schepers, 2005, S. 95). Hier wird also von einer phonologisch existierenden Komponente ausgegangen. gutschrift-Testwörter mit silbenanlautendem h sind beispielsweise ru⟨h⟩ig und dre⟨h⟩en (Valtin, Badel et al., 2003, S. 240). Die Schreibung des ⟨h⟩ in drehen würde nach dem Konzept des SRT silbenstrukturell erfolgen, während in gutschrift als Zugriffsweise (für die Schreibung) vom silbenanlautenden „Laut-h“ ausgegangen wird (Löffler & Meyer-Schepers, 2005, S. 95). Allerdings kann das ⟨h⟩ in den oben genannten Testwörtern ausschließlich bei Überlautung, nicht aber bei Explizitlautung erkannt werden. So ist beispielsweise das ⟨h⟩ in Ruhe nicht durch Explizitlautung [ˈʀuːə], sondern erst durch Überlautung [ˈʀuːhə] ermittelbar. Als Beispielwort führen Löffler und Meyer-Schepers (2005, S. 95) ebenfalls Vieh auf. Um hier das ⟨h⟩ heraushören zu können, müsste zunächst eine Ableitung zum Adjektiv viehisch, gefolgt von dessen Überlautung stattfinden. Damit wird von den gutschrift-Testautorinnen eine lautorientierte Herleitung des silbenanlautenden h angestrebt, die für Schriftlernende

eine Hürde darstellt, da die Überlautung in Anlehnung an das geschriebene Wort, d. h. an die bereits vorhandene orthographisch korrekte Form, geschieht. Durch die Zuordnung des silbeninitialen h bzw. silbenanlautenden h in den Kernbereich oder den erweiterten grammatischen Kompetenzbereich zeigen sich verschiedene Anforderungseinschätzungen zwischen den Testkonzeptionen, auf die im Folgenden weiter eingegangen wird.

### Differenzierung in Schwierigkeits- und Regelbereiche

Der SRT unterscheidet einen Kernbereich von einem Peripheriebereich (Blatt & Frahm, 2013, S. 19). Damit wird der Wortschatz in regelhaft herleitbare Schreibungen, die auf Verstehens- und Wissenstransferbasis gelernt werden können, und Ausnahmeschreibungen eingeteilt (Voss et al., 2007, S. 17 f.). Bei gutschrift findet sich solch eine Einteilung nicht, die z. B. in der AFRA über die Differenzierung in Mehrheits- und Minderheitsschreibungen (vgl. dazu Abschnitt 2.4.3) oder in der Oldenburger Fehleranalyse (OLFA) über Basis- und Orthographeme (G. Thomé & D. Thomé, 2010, S. 9 f.; G. Thomé, 2000, S. 13) erfolgt. Es werden Schwierigkeitsgrade in Form der Niveaubestimmungen bei gutschrift-diagnose angenommen. Die Schreibung von ⟨ie⟩ stellt einen Indikator der erweiterten phonographischen Kompetenz dar. Die Zuordnung zu dieser Kompetenzstufe impliziert eine Schwierigkeitsannahme. Die Verschriftung von beispielsweise ⟨aa⟩ oder Dehnungs-h erfolgen in dem Modell nach derselben Zugriffsweise und demselben Anforderungsniveau, obwohl für die Setzung keine eindeutigen Verschriftungsregularitäten existieren, sondern ausschließlich Hinweise, und die Schreibung von Dehnungs-h in weniger als 20 Prozent der einschlägigen Fälle vorkommt (Bremerich-Vos, 2004, S. 99). Nach Herné (1993, S. 322) stellt sich die Häufigkeitsverteilung der Langvokalschreibungen (ohne /iː/) folgendermaßen dar: 11,6 Prozent mit Dehnungs-h, 0,6 Prozent mit Doppelvokal und 87,8 Prozent ohne Dehnungszeichen. Vokalgraphemverdopplung und Schreibung des Dehnungs-h gelten dementsprechend als Ausnahmeschreibungen im SRT und sind in den Bereich der Peripherie einsortiert, bei dem Wörter überwiegend durch Üben eingeprägt werden müssen (Voss et al., 2007, S. 17 f.). Das Vokalgraphem ⟨ie⟩ wird hingegen innerhalb des SRT dem Kernbereich zugeordnet. Es wird die Regelhaftigkeit bei Wortschreibungen mit ⟨ie⟩ hervorgehoben, da im Geschriebenen die Vokalquantität von /iː/ durch das Graphem selbst angezeigt wird. Naumann und Weinhold (2011, S. 188) erläutern, dass ca. 40 Morpheme mit ⟨i⟩, ⟨ih⟩ oder ⟨ieh⟩ anstelle eines ⟨ie⟩ geschrieben werden. Nach einer Auszählung von G. Thomé (2006, S. 370), in der 10.000 Phonem-Graphem-Relationen aus Texten der deutschen Gegenwartsliteratur analysiert wurden, werden 83 Prozent der Grapheme für /iː/ als ⟨ie⟩ wiedergegeben. Herné und Naumann (2009, S. 11) gehen von einem ähnlichen Mehrheitsverhältnis von 85 Prozent aus.

Eine weitere dreifache Zuordnung ist in der Kategorie „Schreibung von /f/ als ⟨f⟩ und ⟨v⟩“ bei der elementaren grammatischen und phonographisch-silbischen Subskala sowie dem Peripheriebereich erkennbar (vgl. Tabelle 2.11). Bei gutschrift stellen die Schreibungen des Lauts /f/ als ⟨f⟩ und ⟨v⟩ sowie als Vorsilbe in ⟨ver⟩ und ⟨vor⟩ Indikatoren der elementaren grammatischen Teilkompetenz dar. Als Beispielwort für eine Schreibung mit ⟨v⟩ nutzen

Löffler und Meyer-Schepers (2005, S. 95) wieder Vieh. In Lischeid (2006, S. 227) wird ⟨V⟩ieh hingegen unter der elementaren phonographischen Kompetenz betrachtet. Beide Zuweisungen erscheinen auf Basis der von Naumann und Weinhold (2011, S. 188) berichteten Vorkommenshäufigkeit nicht plausibel: ca. 20 Morpheme werden mit ⟨v⟩ statt ⟨f⟩ wiedergegeben. Die Schreibung von ⟨v⟩ wird daher als Minderheitsschreibung bezeichnet. Nach der von Herné (1993, S. 321 f.) dargestellten Häufigkeitsverteilung korrespondieren 11 Prozent der /f/-Grapheme mit ⟨v⟩; 75 Prozent hingegen mit ⟨f⟩ (13 Prozent mit ⟨ff⟩, 1 Prozent mit ⟨ph⟩). Bei dem SRT wird diese Vorkommenshäufigkeit berücksichtigt, indem die Schreibung von z. B. ⟨v⟩ielleicht dem Peripheriebereich zugewiesen wird. Es erfolgt also eine im Vergleich zu gutschrift andere Systematisierung. Schreibungen mit ⟨v⟩ werden dem Peripheriebereich zugewiesen, die Präfixe ⟨ver⟩ und ⟨vor⟩ (wie z. B. in verrenken) hingegen dem Prinzip der Wortbildung und Wörter mit ⟨f⟩ (wie z. B. in Fohlen) dem phonographisch-silbischen Prinzip.

### Rechtschreiberwerb und -entwicklung

Den beiden Tests liegen unterschiedliche Annahmen zu den Erwerbs- und Entwicklungsverläufen des Rechtschreiblernens zugrunde, die damit auch unterschiedliche didaktische Folgerungen und Konsequenzen implizieren. In der Konzeption nach gutschrift ist das normgerechte Schreiben über Fähigkeiten auf verschiedenen Niveaus geregelt, die aufeinander aufbauen und sich im Laufe des Erwerbs der Schriftsprache miteinander verzahnen (Voss, Löffler et al., 2008, S. 134). Zunächst erwerben Schülerinnen und Schüler die elementaren phonographischen und grammatischen Teilkompetenzen, bevor sie die erweiterten erlernen. Es wird von Stufenfolgen mit unterschiedlichen Schwierigkeiten ausgegangen, und diese werden den Klassenstufen in der Schule zugeordnet (Voss, Löffler et al., 2008, S. 134; Löffler & Meyer-Schepers, 2005, S. 84 f.). Dementsprechend sollen Schriftlernende zunächst Laut-Buchstaben-Zuordnungen inklusive einiger Sonderschreibweisen (z. B. Bi- und Trigrapheme) sowie wortgrammatische und morphematische Grundlagen (z. B. elementare Vorsilben, Endungen, Ableitungen, Groß- und Kleinschreibung) in den ersten beiden Jahrgangstufen beherrschen, bevor in den Folgejahrgängen orthografische Regelhaftigkeiten, wie z. B. die Setzung von Kürze- und Längenzeichen, die von den Laut-Buchstaben-Zuordnungen abweichen, erlernt werden (Valtin et al., 2004a, S. 142 f.).

Nach Blatt (2010, S. 113) sollen bei der Erlernung der Schriftsprache zunächst die Wortschreibungen im Kernbereich in drei Lernschritten über das phonologische, das morphologische sowie das Wortbildungsprinzip entdeckend gelernt werden. Hierüber können Schriftlernende die Silbenstruktur im prototypischen Zweisilber untersuchen (phonologisches Prinzip), die vererbten silbenschriftlichen Informationen in flektierten und einsilbigen Formen durch Bildung des prototypischen Zweisilbers aufdecken (morphologisches Prinzip) sowie abgeleitete und zusammengesetzte Wörter analysieren (Wortbildungsprinzip). Auf Grundlage der auf den Prinzipien basierenden Lerninhalte und -verfahren gewinnen die Schülerinnen und Schüler nach Blatt (2010, S. 112) kognitive Einsicht in die Strukturen der Wortschreibung und -bildung. Sie erlangen ein transferfähiges Wissen, das zu einer

selbstständigen Weiterentwicklung der Rechtschreibkompetenz genutzt werden kann und zum Aufbau des Peripheriebereichs benötigt wird. Das Erlernen der Wörter aus diesem Bereich kann auf Basis der vorangegangenen Einsichten und des Wissens gestützt werden (Blatt, 2010, S. 113). Damit grenzt sich der SRT von Modellen mit einem stufenweisen Aufbau ab, indem er die parallele Aneignung verschiedener Prinzipien postuliert (Voss et al., 2007, S. 18).

Entsprechend der beschriebenen verschiedenen Annahmen zum Lernen werden die Schreibprodukte aus entgegengesetzten Blickwinkeln betrachtet. Bei gutschrift erfolgt eine fehleranalytische Auswertung, bei der alle Fehlerarten so vollständig wie möglich erfasst und im Anschluss linguistisch ausgewertet werden (Löffler & Meyer-Schepers, 2005, S. 84). Der SRT versteht sich hingegen als fähigkeitsbasiertes Modell und erfasst die Anzahl richtig geschriebener Indikatoren (Blatt & Frahm, 2013, S. 18).

Die zur fehlerbasierten bzw. fähigkeitsbasierten Auswertung erhobenen Schülerschreibungen werden im Rahmen von gutschrift über ein Lückensatzdiktat erhoben. Als Vorteile dieses Formats kennzeichnen Böhme und Bremerich-Vos (2009, S. 334), dass das unterschiedliche Schreibtempo der Schülerinnen und Schüler kaum ins Gewicht fällt, der Schreibaufwand begrenzt ist, das Schreiben sicher beherrschter Wörter (wie z. B. von Artikeln) entfällt und die Aufmerksamkeit ungeteilt auf der Rechtschreibung liegt, also nicht von Gedächtnisleistungen beeinflusst wird. Das Testformat im SRT unterscheidet sich teilweise, da es zusätzlich jeweils Wert auf einen Testteil mit Sätzen, die vollständig von den Schülerinnen und Schülern zu verschriften sind, legt. In den IGLU-Erhebungen bestand der SRT aus einem reinen Fließtext. Die Diktierzeit (und bei zu digitalisierenden Schreibprodukten die Zeit für die Dateneingabe) erweist sich daher für gutschrift als kürzer. Analysen im Rahmen des SRT haben aber gezeigt, dass gewisse rechtschriftliche Phänomene (wie z. B. die Großschreibung) mit Lückensatzwörtern nicht reliabel und valide erfasst werden konnten (Blatt et al., 2011, S. 250).

gutschrift-diagnose ist ein Modell, das sich an den curricularen Vorgaben orientiert. Die definierten Teilkompetenzen erfassen Fehlerquellen, bzw. umfassen Indikatoren, die auf den Rechtschreibphänomenen fußen, die in den Lehrplänen der Bundesländer beschrieben sind. Zudem weisen die Kompetenzniveaus einen Jahrgangsstufenbezug auf (Valtin et al., 2004a, S. 142 ff.; Valtin, Badel et al., 2003, S. 236 ff.). Daher ist das Konzept voraussichtlich stärker an aktuelle Unterrichtspraxis und -inhalte angepasst bzw. lässt sich einfacher in das Unterrichtsgeschehen integrieren. Der SRT, der auf der Graphematik nach Eisenberg und der didaktischen Umsetzung nach Hinney basiert, und als Bezugseinheit der Schriftsprache die Silbe wählt, bildet Lernwege heraus, die in vielen Rechtschreibunterrichten neu- und andersartig sind.[34] Erfahrungen mit diesem Konzept sind u. a. in Hinney (1997) sowie Hinney und Pagel (2007) beschrieben. Insgesamt lässt sich festhalten, dass sich aus den sprachwissenschaftlichen Hintergründen, die beide Orthografietests charakterisieren, unterschiedliche Vorstellungen von Rechtschreibentwicklung und -lernen ergeben.

[34] Im Rahmen der Lehrerbefragung von IGLU 2006 gaben 1.386 von 4.055 Lehrpersonen (34 Prozent) an, dass im Anfangsunterricht nach dem silbenorientierten Konzept unterrichtet wurde (Mehrfachangaben waren möglich) (Bos et al., 2005, S. 209).

### 2.4.3 Exkurs zu Ausnahmeschreibungen

Der Peripherieberich stellt ein wichtiges Merkmal innerhalb des angenommenen Rechtschreibkompetenzmodells des SRT dar, über den von der Regel abweichende Schreibungen erfasst werden. Er bildet damit das Pendant zum Kernbereich. Im vorherigen Abschnitt erfolgte eine Erläuterung der Analyseeinheiten des Peripheriebereichs und deren Handhabung innerhalb des Konzepts von gutschrift-diagnose. Zum Abschluss dieses Kapitels soll ebenfalls ein Blick auf die in Abschnitt 2.1 dargestellten Rechtschreibkompetenzmodelle der HSP und AFRA gerichtet und der Frage nachgegangen werden, ob und inwiefern dort ebenfalls eine Art Peripheriebereich bestimmt wird. Zunächst wird dabei die HSP und schließlich die AFRA im Zusammenhang mit dem Peripheriebereich des SRT betrachtet.

#### Merkelemente der HSP

Bei der HSP erfolgt eine Unterscheidung zwischen „Regelelementen“ und „Merkelementen“ (May, 2002a, S. 30). Diese Differenzierung wird im Kontext der orthografischen Strategie getroffen. Regelelemente werden beschrieben als solche, „deren Verwendung hergeleitet werden kann“, und Merkelemente als solche, „die sich der Lerner als von der Verschriftung der eigenen Artikulation abweichend merken muss“ (May, 2002a, S. 12). Als Bespiele für die Merkelemente bennent May die folgenden Wörter bzw. Analyseeinheiten: Za⟨h⟩n, ⟨V⟩ater, He⟨x⟩e. Er betont, dass Merkelemente ebenfalls geregelt sind, im Gegensatz zu Regelelementen aber nicht durch bestimmte Verfahren erschlossen werden können (May, 2002a, S. 30). Damit sind die Merkelemente am ehesten mit dem Peripheriebereich des SRT vergleichbar.

Zu den Merkelementen werden von May (2002a, S. 44) fünf verschiedene Rechtschreibphänomene gezählt. Als erstes werden Längenzeichen aufgeführt und konkret die Buchstabenfolgen ⟨ie⟩, ⟨ih⟩, ⟨ah⟩ und ⟨oo⟩ genannt. In der HSP 4/5 werden beispielsweise folgende Lupenstellen mit Dehnungszeichen unter der orthografischen Strategie ausgewertet: G⟨ie⟩ßkanne, Windm⟨üh⟩le, L⟨eh⟩rerin. May (2002a, S. 30) erläutert, dass es zu den Längenzeichen zwar allgemeine Regeln gibt, doch für die Kennzeichnung der Länge nur Wahrscheinlichkeitsbeziehungen gelten. Im Gegensatz dazu beschreibt er die Kürzezeichen als analytisch operativ erschließbar (May, 2002a, S. 44). Im SRT stellen die genannten Lupenstellen, bis auf die Verschriftung des lang gesprochenen /iː/, ebenfalls Struktureinheiten des Peripheriebereichs dar. Das Vokalgraphem ⟨ie⟩ wird, wie in Abschnitt 2.4.2 dargestellt, unter dem Kernbereich ausgewertet.

Als zweites unter die Merkelemente fallende Phänomen benennt May (2002a, S. 16) die Schreibung von ⟨x⟩ (wie z. B. bei Max), da dieses lautlich über mehrere Grapheme (⟨chs⟩, ⟨ks⟩, ⟨cks⟩ und ⟨gs⟩) repräsentierbar ist. Zum SRT finden sich in der bisherigen Literatur keine konkreten Zuweisungen von Testwörtern mit dem Graphem ⟨x⟩. Eisenberg (2006a, S. 306 f.) ordnet es nicht dem Kernbestand der Grapheme, sondern nur einem erweiterten Grapheminventar des Deutschen, zu. Er schreibt, dass der Buchstabe hauptsächlich in Fremdwörtern Anwendung findet und im Kernwortschatz als besondere Schreibung gilt

(Eisenberg, 2011, S. 86; Eisenberg, 2006c, S. 70). Wie der Buchstabe im SRT ausgewertet würde, kann an dieser Stelle zwar vermutet, aber nicht eindeutig beurteilt werden.

Die Schreibungen von ⟨qu⟩ und ⟨v⟩ (nicht jedoch in Präfixen) gelten als weitere Merkelemente in der HSP. Die Affrikate ⟨qu⟩ wird hinzugezählt, da sie alphabetisch als ⟨kw⟩ darstellbar ist. Beide Grapheme sind auch Items des Peripheriebereichs, wie zuvor in Abschnitt 2.4.1 im Vergleich mit gutschrift-diagnose dargelegt wurde. Als letztes Merkelement wird von May (2002a, S. 44) ⟨ß⟩ als Zeichen für das stimmlose /s/ aufgeführt, „wie in Gießkanne, da das gleiche Phonem alphabetisch auch durch ⟨s⟩ bezeichnet werden kann (Gras)“. Das Graphem ⟨ß⟩ wird beim SRT im Kernbereich betrachtet: Wörter in der Grundform sind dem phonographisch-silbischen und in flektierter und abgeleiteter Form dem morphologischen Prinzip zugeordnet.

### Minderheitsschreibungen der AFRA

Das Gliederungsprinzip der AFRA in Mehr- und Minderheitsschreibungen, das auf sprachstatistischen Häufigkeitsverteilungen basiert, weist ebenfalls eine Reihe von Ähnlichkeiten mit dem Peripheriebereich des SRT auf. Minderheitsschreibungen werden von Herné (2003, S. 893) als „Ausnahmefälle“ bezeichnet. Dabei handelt es sich um Schreibungen, die von den Prinzipien der deutschen Orthografie abweichen sowie um Allographe, die nicht den Regelfall, sondern die Ausnahme oder Minderheit bilden (Herné & Naumann, 2009, S. 8; Herné, 1993, S. 323). Sie sollen nun aufgeführt (vgl. im Folgenden jeweils Tabelle 2.2) und ihre Verortung im Kompetenzmodell des SRT dargestellt werden.

Unter der ersten AFRA-Fehlerebene, der Phonem-Graphem-Korrespondenz, werden als Minderheitsschreibungen die Kategorien spezielle Grapheme (SG-) und spezielle Verbindungen (SV-) gefasst. Bei SG- wird die Schreibung von ⟨v⟩ für /f/ (davon unabhängig: Präfixe) sowie von ⟨ß⟩ genannt. Das Graphem ⟨v⟩ stellt ebenfalls eine Struktureinheit des Peripheriebereichs dar. Das ⟨ß⟩ wird hingegen im Kernbereich betrachtet (s. o.). Unter der Kategorie SV- führen Herné und Naumann die Verbindungen ⟨ai⟩, ⟨chs⟩, ⟨pf⟩ und ⟨qu⟩ auf. Dass ⟨ai⟩ gleichermaßen eine Indikatorstelle des Peripheriebereich wäre, erscheint plausibel, kann aber aus der bisher vorliegenden Literatur zum SRT nicht eindeutig entnommen werden. Die Schreibungen von ⟨chs⟩ und ⟨qu⟩ besitzen im SRT-Konzept auch eine Ausnahmestellung. Wörter mit diesen Buchstabenfolgen sind auf Basis der SRT-Indikatorzuweisungen in dieser Arbeit als Analyseeinheiten des Peripheriebereichs kodiert und ausgewertet worden. Die Affrikate ⟨pf⟩ wird im Peripherie- als auch im Kernbereich betrachtet. Als Silbengelenk stellt sie ein Item des phonographisch-silbischen und (bei Flexion und Derivation) des morphologischen Prinzips sowie im Silbenanfangsrand eines des Peripheriebereichs dar.

Die zweite Fehlerebene in der AFRA ist die Vokalquantität. Die dort aufgeführten Minderheitenschreibungen langes i (LI-), lange Vokale (LV-) und Kurzvokale (KV-) sind jeweils Analyseeinheiten des Peripheriebereichs. In diesem Zusammenhang wurden sie zum Teil bereits ausführlich innerhalb des konzeptionellen Vergleichs mit gutschrift dargestellt (vgl. Abschnitt 2.4.2), weshalb nun nur kurz darauf eingegangen wird. Die Analyse der Schrei-

bung von /i:/, die nicht mit ⟨ie⟩ wiedergegeben wird, erfolgt in der AFRA-Systematik unter der Kategorie LI-. Im SRT fallen diese Schreibungen (mit ⟨i⟩, ⟨ih⟩ oder ⟨ieh⟩) analog unter den Peripheriebereich; ebenso Wörter mit Dehnungs-h und Doppelvokalbuchstaben (LV-), wie auch irreguläre Kurzvokalschreibung in Wörtern wie z. B. an (*ann) oder Batterie (*Baterie).

Die Morphologie stellt eine weitere Ebene der AFRA dar. Abweichungen der Auslautverhärtung werden unter der Kategorie konsonantische Ableitungen (KA-) ausgewertet (wie u. a. die Schreibung von Jugend: da das Wort keinen Plural hat, kann der konsonantische Auslaut nicht abgeleitet werden). Unter die vokalische Ableitung (VA-) erfolgt eine Sammlung zu Ausnahmen bei der Ableitung der Umlaute. Wörter wie Eltern (statt *Ältern) oder Säule (statt *Seule) fallen darunter. Diese nicht regelhaft herleitbaren Fälle sind in dem Konzept des SRT ebenfalls unter dem Bereich der Peripherie verortet.

Die letzte Ebene ist die der Syntax. Hier wird die Kleinschreibung als Normalfall und die Großschreibung als Sonderfall angesehen. Unter die Minderheitsschreibungen zählen die Großschreibung am Satzanfang, der Höflichkeitsanrede, von Eigennamen, von Konkreta und Abstrakta sowie von Substantivierungen (Herné & Naumann, 2009, S. 16). Die Groß- und Kleinschreibung wird im SRT unter dem wortübergreifenden Prinzip betrachtet. Es erfolgt keine Differenzierung in unterschiedliche Subskalen auf Basis des Grads der Anforderung und Schwierigkeit.

### Resümee

Die Ausführungen zeigen, dass es im Hinblick auf das, was von den Testautoren von HSP, AFRA und SRT als Ausnahme oder Minderheit im Bereich der Wortschreibung angesehen wird, viele Überschneidungen gibt. Zwischen den Merkelementen der HSP, den Minderheitenschreibungen der AFRA und dem Peripheriebereich des SRT zeigen sich eine Reihe von Gemeinsamkeiten. Alle in Tabelle 2.11 gelisteten Rechtschreibphänomene, die dem Peripheriebereich zugehörig sind, werden bei der HSP ebenfalls als Merkelemente und bei der AFRA als Minderheiten thematisiert. Umgekehrt gilt dies nicht gleichermaßen. Differenzen zeigen sich bei der HSP bei der Schreibung von ⟨ie⟩ und bei der AFRA von ⟨ß⟩, teilweise von ⟨pf⟩ sowie bei der Groß- und Kleinschreibung. Damit werden mehr rechtschriftliche Phänomene im Konzept des SRT als regelhaft herleitbare Schreibungen klassifiziert. Der hier vorgenommene und auf den Peripheriebereich des SRT abzielende Vergleich legt einen Ausschnitt an Gemeinsamkeiten und Unterschieden zwischen den verschiedenen Testkonzeptionen offen. Ein vollständiger theoretischer Vergleich, der alle Strategiebereiche, Fehlerebenen und -kategorien sowie Prinzipien mit ihren Items berücksichtigt, geht über die Zielsetzung dieser Arbeit hinaus und sollte in einem zukünftigen Forschungsvorhaben angestrebt werden.

# 3 Anlage und statistische Auswertungsmethoden

Die mit der vorliegenden Arbeit vorgestellte IGLU-Zusatzstudie Rechtschreibung (ZuRecht) wurde unter der Leitung von Wilfried Bos und Andreas Voss konzipiert und gemeinsam mit der Autorin durchgeführt. In diesem Kapitel werden die Anlage der Studie und die methodischen Verfahren der Datenauswertung erläutert. Konkret erfolgt zunächst eine Darstellung der Einbindung des Projekts in IGLU, der Untersuchungsgruppe, des Erhebungsdesigns sowie der orthografischen Testinstrumente (Abschnitt 3.1). Im Anschluss werden der Prozess und die Regeln der systematischen Dateneingabe der Schülertexte sowie der kompetenzmodellorientierten Kodierung der Schreibprodukte beleuchtet (Abschnitt 3.2). Vervollständigt wird das Kapitel durch eine Erläuterung der Datenauswertungsverfahren, die die Grundlage für die Ergebnisdarstellungen in Kapitel 4 bilden und daher ausführlich behandelt werden (Abschnitt 3.3).

## 3.1 Erhebungsrahmen

ZuRecht bettet sich in die im Rahmen von IGLU durchgeführten Erhebungen ein. Bereits bei der ersten IGLU-Erhebung 2001 wurde die Rechtschreibkompetenz der Schülerinnen und Schüler mit DoSE erfasst. Es nahmen allerdings nicht alle Bundesländer an IGLU-E Orthografie teil: Aus 12 Ländern liegen Daten von 3.391 Kindern vor. Infolge der unerwarteten und unbefriedigenden Leistungsergebnisse (vgl. Abschnitt 2.2.2) wurde der Rechtschreibung in IGLU eine verstärkte Aufmerksamkeit gewidmet. Als Vorbereitung für die zweite IGLU-Erhebung wurde ein neuer Rechtschreibtest konzipiert und pilotiert: der SRT. Die Erprobung erfolgte in der Voruntersuchung zu IGLU 2006 (vgl. Abschnitt 2.3.2). Im Anschluss wurde er für den Einsatz in der Haupterhebung überarbeitet und gemeinsam mit der gutschrift-diagnose sowie der HSP in einem rotierten Design eingesetzt. Für diese Erhebung liegt eine bundesweite Stichprobe von rund 8.000 Schülerinnen und Schülern vor.

Um ZuRecht an die IGLU-Hauptuntersuchung anzuschließen, erfolgte die Durchführung nach vergleichbaren Bedingungen. Es wurden ebenfalls gutschrift-diagnose, der SRT, die HSP sowie ein weiterer Rechtschreibtest, die *Münsteraner Rechtschreibanalyse* (MRA) von Friedrich Schönweiss, eingesetzt.[1] In der vorliegenden Arbeit werden der gutschrift-Test und der SRT analysiert sowie vergleichend gegenübergestellt, da beide Tests auf linguistisch basierten Kompetenzmodellen fußen (z. B. im Gegensatz zur HSP, der ein „lern- und entwicklungspsychologisches Konzept“ zugrunde liegt (May & Malitzky, 1999, S. 7)). Die Erhebung fand ebenfalls am Ende der Primarstufe und im Klassenverband statt.

Die Teilnahme war freiwillig; es konnten elf Dortmunder Grundschulen mit 26 Klassen für die Studie gewonnen werden. Der Großteil der Schulen ist zweizügig; zwei Schulen haben vier vierte Jahrgangsstufen. Für die Testdurchführung wurden zumeist mehrere Testleiter parallel eingesetzt. Andernfalls wurden die Klassen innerhalb eines Tages konsekutiv besucht, um den ersten bzw. zweiten Teil der Erhebung am selben Tag in einer Schule abzuschließen. Der Zeitumfang für die Bearbeitung der Rechtschreibtests betrug jeweils rund 20 Minuten. Um die Belastung der Schülerinnen und Schüler möglichst gering zu halten und um Motivationsnachlass sowie Ermüdungserscheinungen vorzubeugen, wurde die Erhebung auf zwei Testtage aufgeteilt, wobei die Rechtschreibtests untereinander in der Abfolge gemischt wurden, um Reihenfolgeeffekte auszuschließen. Für die beiden Testtage wurde jeweils eine Zeitstunde reserviert. Zwischen den Tests gab es eine kleine Pause zur Entspannung, in denen die Schülerinnen und Schüler z. B. den Stift zur Seite legen und die Hände ausschütteln sollten. Im Anschluss an die Bearbeitung des letzten Rechtschreibtests wurde ein Kurzfragebogen administriert, um einige Hintergrundinformationen zu den Kindern zu gewinnen (vgl. Abschnitt 4.3). Die in dieser Arbeit berichteten Daten sind aus Gründen des Datenschutzes in der Form anonymisiert, dass keine Namen von beteiligten Personen oder Institutionen genannt, sondern ausschließlich über numerische Schlüssel voneinander unterschieden und betitelt werden.

Die Anzahl der Schülerinnen und Schüler, die gutschrift und/oder den SRT bearbeitet haben, beläuft sich auf insgesamt 581. Tabelle 3.1 schlüsselt die Stichprobengröße getrennt nach Orthografietest auf. Hier ist zu erkennen, dass 566 Schülerinnen und Schüler den gutschrift-Test, 19 davon aber nicht den SRT geschrieben haben. Wiederum haben 562 Kinder am SRT teilgenommen, aber 15 davon nicht an gutschrift. Insgesamt liegen damit für 547 Kinder Daten für beide Rechtschreibtests vor. Den gutschrift-Test und den SRT haben also 94 Prozent der Schülerinnen und Schüler bearbeitet. Der simultane Einsatz zweier Rechtschreibtests mit identischer Stichprobe wurde als integraler Bestandteil des Studiendesigns gewählt, da er unmittelbar eine empirisch vergleichende Analyse der Tests ermöglicht.

Die Durchführung der Testsitzung erfolgte auf Basis eines umfangreichen Testleitermanuals und -skripts, in dem der Ablauf und die Instruktionen im Einzelnen beschrieben standen,

[1] Da nach mehrfachen Anfragen beim Testautor für die MRA kein differenzielles und transparentes Kompetenzmodell mit einem Kategoriensystem für die Kodierung offengelegt wurde, konnten diese Daten bisher nur auf der Ebene ganzer Wörter ausgewertet werden. Für die HSP liegen ebenfalls globale Analyseergebnisse zur Gesamtstichprobe vor (Kowalski & Voss, 2009).

| | | gutschrift | |
|---|---|---|---|
| | | 19 | |
| SRT | 15 | **547** | 562 |
| | | 566 | |

Tabelle 3.1: Stichprobengröße der bearbeiteten Rechtschreibtests

um eine möglichst identische und standardisierte Erhebungssituation in allen Klassen zu gewährleisten.[2] Die Instruktionen und Texte wurden aus der IGLU-Hauptuntersuchung adaptiert, sodass die üblichen Vorgaben und Regularitäten nach dem Vorbild einer Testsitzung in Large-Scale-Assessments gewahrt wurden (z. B. Verhindern von Abschreiben, keine Beantwortung von Rechtschreibfragen, Anwesenheit einer Lehrperson etc.). In den spezifischen Anweisungen für die Durchführung der Orthografietests war festgelegt, dass die Wörter deutlich, aber in keiner speziellen Betonung oder Hervorhebung von Silben oder anderen Wortbestandteilen vorgelesen werden sollten. Diktiert werden sollte zunächst der ganze Satz und im Anschluss die Lückenwörter (gutschrift) bzw. sinnhafte, vorgegebene Satzabschnitte (SRT), die die Kinder dann mitschreiben sollten. Bei dem SRT wurden die Testleiterinnen und Testleiter angewiesen, auch die Satzzeichen zu diktieren (im gutschrift-Test sind keine von den Kindern zu setzenden Interpunktionszeichen vorhanden). Am Ende erfolgte noch einmal eine vollständige Wiederholung aller Sätze.

### Die Rechtschreibtexte

Bei gutschrift wurde als Testinstrument zur Erfassung der Rechtschreibkompetenz ein Lückentext konzipiert, in den fehlende Wörter einzutragen sind. Das Lückensatzdiktat aus ZuRecht ist in Abbildung 3.1 dargestellt.[3] Die Testwörter, die die Kinder schreiben sollten, sind bei den ersten vier Sätzen unterstrichen und im Weiteren chronologisch aufgeführt. Der Test besteht aus insgesamt zwölf Sätzen und 35 Testwörtern. Die Sätze sind inhaltlich nicht zusammenhängend und thematisieren unterschiedliche Alltagssituationen. Mit dem Test werden sowohl die elementaren als auch die erweiterten Teilkompetenzen erhoben (Löffler & Meyer-Schepers, 2009, S. 64).[4]

Im SRT wurde die orthografische Kompetenz mit einem zusammenhängenden Text ermittelt, der vollständig diktiert wurde (vgl. Abbildung 3.2) (Blatt & Jarsinski, 2009, S. 99). Der Fließtext wurde in identischer Form in ZuRecht und in IGLU-E eingesetzt. Er fußt auf der IGLU-E-Voruntersuchung und wurde auf Basis der dortigen Ergebnisse optimiert.

[2] Die Autorin der vorliegenden Arbeit war im Rahmen von ZuRecht für die Koordination der Erhebungen an den Schulen und für die Schulung der Testleiter zuständig sowie selbst als Testleiterin im Einsatz.

[3] Aus urheberrechtlichen Gründen sind hier nur die ersten elf Testwörter in Sätze eingebunden und die weiteren anschließend separat aufgelistet (Löffler & Meyer-Schepers, 2009, S. 64, 67).

[4] Die Tests gutschrift-1 und gutschrift-2 prüfen hingegen ausschließlich die Teilkompetenzen auf elementarer Stufe ab (Löffler & Meyer-Schepers, 2009, S. 64; Voss, Löffler et al., 2008, S. 136).

Die Zeitschrift mit dem Fernsehprogramm liegt auf der Eckbank.
Kommt heute mit euren Fahrrädern bei mir vorbei. Dann machen wir ein Wettrennen, wer von uns der Schnellste ist.
Alle lachen, weil Marie wegen ihrer Zahnspange mit der Zungenspitze an den Vorderzähnen anstößt.

| | | |
|---|---|---|
| Frühstück | vorbereitet | räumt |
| schließlich | Geschirr | Spülmaschine |
| Sonnenstrahlen | viele | Hautschäden |
| Sonnenschutzmittel | ölig | informieren |
| verbrennt | Kälte | fließt |
| Läufern | Nahrung | Spaziergangs |
| verdeckt | geblitzt | ruhig |
| Vorsicht | empfindlich | aufgepasst |

Abbildung 3.1: Das Lückensatzdiktat des gutschrift-Tests

Die Zauberwiese
Ein älteres Pferd fragt ein neugieriges Fohlen: „Willst du vielleicht bunte Stifte mit Klettband?“ Das Fohlen, das eine riesengroße Freude am Malen hat, nickt begeistert. Es nimmt die schicken Farbstifte mit dem linken Vorderhuf. Es beginnt schließlich die Wiese zu schmücken. Bald sieht man herrliche Bilder auf dem Gras. Ein scheues Reh grast friedlich am Fluss. Ein listiger Fuchs schielt nach einer weißen Gans. Ein Kätzchen mit dünnen Schnurrbarthaaren schläft ein bisschen. Zwei leuchtende Schwäne verrenken ihre Hälse. In der Mittagssonne glänzen und glitzern die Gemälde. Da kommen süße Frösche angehüpft. Sie begrüßen das glückliche Fohlen. Sie quaken vor Spaß und Vergnügen. Alle Pferde auf der Weide schütteln fröhlich ihre Mähnen. Sie lassen den Künstler mit lautem Wiehern hochleben.

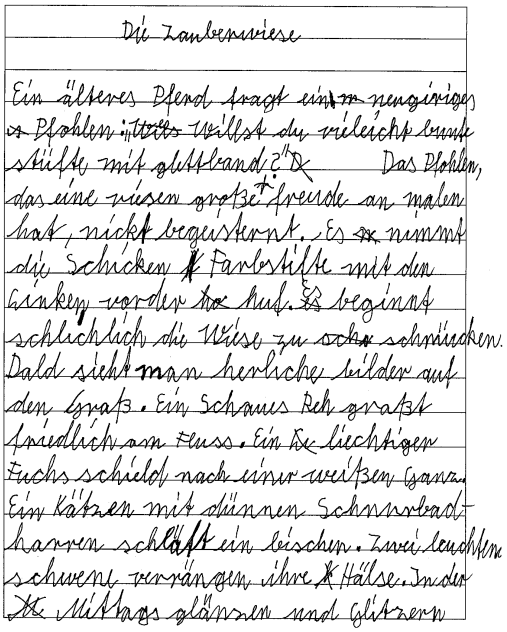
Abbildung 3.2: Das Fließtextdiktat des SRT

Bei dem Text handelt es sich um eine fantastische Geschichte, die in der Natur spielt und in der verschiedene Tiere miteinander interagieren. Das Diktat setzt sich aus 15 Sätzen, inklusive Nebensätzen und Sätzen in wörtlicher Rede, mit 121 Wörtern zusammen.

## 3.2 Transkribierung und Kodierung

### Transkribierung

Ziel der Dateneingabe der Schülerdiktate war eine möglichst exakte 1:1-Übertragung der Schülerschreibungen in eine digitale Version. Ein Beispiel für einen Schülertext ist in Abbildung 3.3a dargestellt. Die entsprechende transkribierte Version zeigt Abbildung 3.3b.



(a) Schülertext

#w001Die #w002Zauberwiese

#w003Ein #w004älteres #w005Pferd #w006fragt #w007ein #w008neugiriges
#w009Pfohlen: #w010„Willst #w011du #w012vieleicht #w013bunte
#w014stüfte #w015mit #w016glettband?“ #w017Das #w018Pfohlen,
#w019das #w020eine #w021riesen_große #w022freude #w023an #w024malen
#w025hat, #w026nickt #w027begeistert. #w028Es #w029nimmt
#w030die #w031Schicken #w032Farbstifte #w033mit #w034den
#w035Linken #w036vorder_huf. #w037Es #w038beginnt
#w039schlichlich #w040die #w041Wiese #w042zu #w043schmücken.
#w044Dald #w045sieht #w046man #w047herliche #w048bilder #w049auf
#w050den #w051Graß. #w052Ein #w053Schaues #w054Reh #w055graßt
#w056friedlich #w057am #w058Fluss. #w059Ein #w060liechtiger
#w061Fuchs #w062schield #w063nach #w064einer #w065weißen #w066Ganz.
#w067Ein #w068Kätzen #w069mit #w070dünnen #w071Schnurbad-
harren #w072schläft #w073ein #w074bischen. #w075Zwei #w076leuchtene
#w077schwene #w078verrängen #w079ihre #w080Hälse. #w081In #w082der
#w083Mittags #w084glänzen #w085und #w086Glitzern

(b) Transkribierung

Abbildung 3.3: Beispiel für ein Schülerdiktat (Ausschnitt)

Die Dateneingabe koordinierte und konzipierte die Autorin und erfolgte von ihr zusammen mit Studierenden des Lehramtes und der Erziehungswissenschaften. Die Anweisungen wurden ausführlich in einem Manual mit Beispielmaterial zur Veranschaulichung beschrieben. Die Regeln zur Dateneingabe entstammen dabei u. a. aus den Erfahrungen, die die Autorin im Rahmen der Dateneingabe bei der IGLU-Voruntersuchung gewonnen hatte, und wurden für ZuRecht weiterentwickelt und optimiert.

So wurde eine Art *Vier-Augen-Prinzip* etabliert: Jeweils zwei sich abwechselnde Personen waren dabei für die Transkribierung eines Schülertextes zuständig und standen bei Zweifelsfällen im Austausch miteinander bzw. mit der Autorin dieser Arbeit. Die erste Person hatte die Aufgabe, die getätigten Schreibfehler eines Kindes in dessen Text zu markieren. Im Anschluss wurde der Schülertext mit den markierten Fehlern von einer zweiten Person durchgesehen und kontrolliert sowie die Fehler in eine Eingabemaske übertragen. Bei der Eingabemaske handelt es sich um eine Vorlage mit dem fehlerfreien Lücken- bzw. Diktattext. Die Schülertexte wurden also nicht mehr (wie in der Voruntersuchung) abgetippt, sondern die Schreibfehler in eine Vorlage übertragen. Dieses Vier-Augen-Prinzip erwies sich als sehr präzise, wie Kontrollen der Dateneingabe zeigten. Die Vorlage zur Dateneingabe bzw. Fehlerübertragung liegt als reines Textformat vor. Sie beinhaltete u. a. ein Feld zur Übertragung der anonymisierten Schüleridentifikationsnummern.

Zudem wurden allen Wörtern der Tests Wortidentifikationsnummern zugewiesen. Dabei handelt es sich um laufende Nummern innerhalb eines Tests. Über diese Wort-IDs kann sichergestellt werden, dass bei der anschließenden Datenauswertung genau das vom Kind

vorgesehene Wort kodiert wird. Aufgrund von wiederholt auftretenden und ähnlichen Wörtern sowie unleserlichen Schreibungen sind Wort-IDs notwendig. Beispielsweise kommt das Testwort Fohlen im Diktat des SRT dreimal vor. Ebenfalls sind u. a. Artikel und Pronomen mehrmals auftretende Testwörter. Bei der weiteren Datenverarbeitung bleibt über die Wort-IDs die Information der Reihenfolge des Auftretens erhalten. Schließlich gibt es Wortschreibungen, die stark von der normgerechten kodifizierten Form abweichen. Die Wort-IDs waren ebenfalls hilfreich, um diese eindeutig zuzuordnen.

Bei der Dateneingabe wurde das individuelle Schriftbild der Schülerinnen und Schüler berücksichtigt, um Zweifelsfälle bei der Entzifferung einzelner Buchstaben oder Zeichen aufzulösen. Wie eingangs erwähnt, soll die digitale Version der Schreibungen ein möglichst genaues Abbild der originalen Papierversion darstellen. Daher wurden z. B. Interpunktions- und Oberzeichenfehler, falsche Zusammen- und Auseinanderschreibungen sowie ausgelassene und unleserliche Buchstaben und Wörter berücksichtigt. Fehlende und unleserliche Wörter bzw. Buchstaben wurden beispielsweise über spezifische Zeichenfolgen als entsprechende Platzhalter eingegeben. Hat z. B. ein Kind ein Wort nicht geschrieben, wurde dreimal ein „x“ eingegeben. Falls ein Kind ein Wort fälschlicherweise auseinandergeschrieben hat, wie beispielsweise das Wort riesengroße, wurde dies durch einen Unterstrich, also riesen_große, transkribiert (vgl. Abbildung 3.3). Die Qualität der Dateneingabe wurde regelmäßig von der Autorin stichprobenartig kontrolliert. Mit den genauen und ausführlichen Anweisungen zur Dateneingabe und den Kontrollfunktionen konnte eine qualitativ hochwertige, standardisierte und objektive Transkribierung sichergestellt werden.

### Kodierung

Die Schülertests, die in eine digital aufbereitete Form überführt wurden, sind im Anschluss einer quantitativen Inhaltsanalyse unterzogen worden. Unter dem Begriff sammeln sich eine Reihe von Verfahren, die jeweils ihr spezielles Vorgehen und ihre eigenen Schwerpunkte haben (Bos & Tarnai, 1989, S. 1). Die hier verwendete Frequenzanalyse beschränkt sich auf die Auszählung und den Vergleich manifester Texteinheiten (Bos & Tarnai, 1989, S. 4). Dabei werden die Schreibprodukte der Schülerinnen und Schüler den ausgewählten Indikatoren zur Messung der Teilkompetenzen zugeordnet und kodiert. Die Kodierung ist ein Vorgang, bei dem Itemantworten von befragten Personen (hier die Schreibungen von Worteinzelbestandteilen der Kinder) in eine numerische Darstellung transformiert werden, damit sie mit einem Testmodell ausgewertet werden können (Rost, 2004, S. 78). Dies geschieht für den gutschrift-Test und den SRT auf unterschiedlichem Wege und soll im Folgenden dargestellt werden.

Die eingegebenen Schülertests werden bei gutschrift-diagnose über ein eigens entwickeltes Computerprogramm ausgewertet. Die Testautorinnen bezeichnen das System als ein *Expertensystem* (XPS), mit dem die Schreibungen der Schülerinnen und Schüler nach dem Konzept des Kompetenzmodells kodiert werden. Um die elektronische Datenverarbeitung zu ermöglichen, haben Löffler und Meyer-Schepers linguistische Fehlerkategorisierungen vorgenommen. Dies bedeutet, dass die getätigten Fehlschreibungen und ihre Zuordnung zu

den Kompetenzdimensionen und -stufen in das System eingepflegt werden (Voss, Löffler et al., 2008, S. 136 f.). Sie basieren auf dem Einsatz in vorherigen Studien, wie z. B. IGLU 2001. Das System wurde auch durch Kodierungsvorschriften ergänzt, da neue Einzelfehler von den im Rahmen von ZuRecht getesteten Kindern produziert wurden, die das Programm bisher noch nicht erfasst hatte. Durch den Einsatz des XPS ist eine Auswertung großer Datenmengen zeitnah möglich (Voss, Löffler et al., 2008, S. 137).

Die Ergebnisse der computergestützten Kodierung wurden in Tabellenform ausgegeben. Sie beinhalten die Vergabe von abgestuften Punktwerten für falsche, teilweise falsche und nicht falsche Schreibungen. Die Information, welcher Indikator innerhalb eines Wortes analysiert wird, wurde nicht von den gutschrift-Autorinnen transparent gemacht. Im Anschluss wurden die Punktewerte von der Autorin dieser Arbeit in ein statistisch auswertbares Format transformiert, bei dem pro Zeile ein Kind mit seinen Testresultaten in numerischen Werten aufgeführt ist. Diese Datenmatrix wurde daraufhin weiterverarbeitet. Die Schritte des Datenaufbereitungs- und Analyseprozesses sind in Abschnitt 4.2.1 beschrieben. Das XPS berücksichtigt Wörter, die die Schülerinnen und Schüler ausgelassen haben. Diese wurden entsprechend als Missings kodiert und in den Analysen als fehlende Werte behandelt. Der Anteil an nicht geschriebenen Wörtern fällt gering aus. Theoretisch kann bei der Stichprobengröße von 566 Schülerinnen und Schülern sowie 35 Testwörtern von insgesamt 19.810 geschriebenen Wörtern ausgegangen werden. Der Anteil an Missings beläuft sich auf 0,07 Prozent, da 13 Wörter fehlend sind.

Der SRT wurde mit einer im Rahmen von NEPS entwickelten, computerbasierten Kodierungssoftware ausgewertet (Frahm, 2012). Es handelt sich dabei um den *SRT-Editor*, der in Kooperation zwischen der Universität Hamburg (Inge Blatt und Sarah Frahm) und dem Deutschen Institut für Internationale Pädagogische Forschung (DIPF) (Ulf Kröhne und Thomas Martens) entstanden ist (Strietholt et al., 2013, S. 574; Frahm et al., 2011, S. 226). Anders als bei dem Vorgehen in der IGLU-Voruntersuchung unter Verwendung von MaxQda (vgl. Abschnitt 2.3.2) wurden hier von der Autorin der vorliegenden Arbeit nicht die Schreibungen der Kinder im Einzelnen betrachtet, sondern Regeln für jedes Testwort direkt in den SRT-Editor eingepflegt. Die Kodierungsregeln wurden von Blatt vorgegeben. Da der SRT-Editor die Regeln unabhängig von den Schülertexten abspeichert, können im Anschluss auch beliebige weitere Stichproben mit dem einmal erstellten Regelwert kodiert werden. Die Kodierung kann dann automatisch erfolgen. Die Ergebnisse der Kodierungen mit MaxQda konnte die Autorin mit dem SRT-Editor reproduzieren. Mit beiden Kodierungsverfahren konnten nahezu identische Resultate erzeugt werden. Informationen zu den Regelbildungen zur Auswertung der Struktureinheiten innerhalb der Software sind in Frahm (2012, S. 127 ff.) beschrieben.

Der SRT-Editor erstellt als Ergebnis des Kodierungsprozesses eine Datenmatrix, die sofort, z. B. mit der Statistiksoftware „IBM SPSS Statistics“, weiterverarbeitet werden kann. Die Daten liegen hier, im Gegensatz zu gutschrift, in einem dichotomen Format vor. Schreibungen von Struktureinheiten wurden also binär mit falsch oder richtig bewertet. Zum Zeitpunkt der Datenauswertung in ZuRecht konnten Missings bei der Kodierung nicht berücksichtigt werden und wurden als Fehler bzw. falsche Schreibweisen markiert.

Mittlerweile füllt das Programm auch diese Lücke. Die Anzahl an nicht geschriebenen Wörtern, die für die differenzielle Datenauswertung[5] benötigt werden, beläuft sich auf 213. Die theoretische Menge an Wortschreibungen liegt bei 43.836 Wörtern (562 Schülerinnen und Schüler sowie 78 Testwörter). Damit hat der SRT eine Missingquote von 0,49 Prozent.

## 3.3 Probabilistische Testverfahren

In diesem Kapitel werden die wesentlichen Eigenschaften der Verfahren dargestellt, die Anwendung bei der Datenauswertung in Kapitel 4 finden. Zunächst sollen hierfür in Abschnitt 3.3.1 grundlegende Begriffe und Merkmale probabilistischer Testmodelle beschrieben sowie durch einen Vergleich mit der klassischen Testtheorie herausgestellt werden. Im Anschluss erfolgt in Abschnitt 3.3.2 eine testtheoretische Beschreibung des Raschmodells. Dabei wird ausführlich auf die Grundgleichung, den daraus abgeleiteten Funktionsverlauf sowie die notwendigen Eigenschaften raschkonformer Tests eingegangen. Erläuterungen zur mehrdimensionalen Skalierung und Parameterschätzung des Raschmodells erfolgen im Anschluss. Die Abschnitte 3.3.2.1 und 3.3.2.2 behandeln Mechanismen und Werkzeuge der Modellwahl via Modellgeltungstests und der Testoptimierung via Itemanalyse, die ebenfalls für die Analysen in Kapitel 4, bei denen konkurrierende Modelle gegenübergestellt werden, grundlegend sind. Im darauffolgenden Abschnitt 3.3.3 wird ein weiteres probabilistisches Testverfahren vorgestellt: die latente Klassenanalyse. Auch hier erfolgt eine Einführung in die Modellgleichung und die Grundannahmen sowie darauffolgend ein kurzer Überblick über ihr Verhältnis zum Raschmodell. Wahl- und Gütekriterien für probabilistische Clustermodelle sind ferner erläutert, bevor ein kleiner Exkurs über den Einsatz der latenten Profilanalyse am Beispiel der großen Schulleistungsstudie IGLU Abschnitt 3.3 abrundet.

### 3.3.1 Merkmale und Abgrenzung

#### Latente Variable/Konstrukt

Für die statistische Auswertung der Testdaten wurde ein probabilistisches Testverfahren ausgewählt. Probabilistische Testtheorien (PTT) werden auch *Item-Response-Theorien* (IRT) genannt und bezeichnen spezifische Testmodelle. Ziel der Modelle ist es, systematische Zusammenhänge zwischen den Antworten von Personen auf Testaufgaben oder -items über latente Personenfähigkeiten oder -eigenschaften aufzudecken (Rost, 2004, S. 28 f.). Bei latenten Personenfähigkeiten handelt es sich zumeist um komplexe Merkmale einer Person, wie beispielsweise Intelligenz oder Lernmotivation, die nicht direkt beobachtbar

[5] Die Struktureinheiten, die für die Auswertung nach den fünf Teilkompetenzen verwendet werden, umfassen nicht alle Wörter des Diktates, das insgesamt 121 Wörter beinhaltet. Beispielsweise weisen viele kleine Wörter, wie u. a. Artikel oder Konjunktionen, keine oder kaum Analyseeinheiten auf.

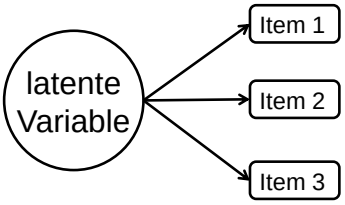



Abbildung 3.4: Zusammenhang von latenter Personenvariable und manifesten Variablen (in Anlehnung an Rost, 2004, S. 30)

sind (Gniewosz, 2011b, S. 68). Um dieses Problem zu lösen, werden in der IRT zwei Ebenen von Variablen unterschieden: latente und manifeste Variablen.

Latente Variablen bezeichnen die latenten Eigenschafts- oder Merkmalsausprägungen einer Person (Molenaar, 1995, S. 3). Mit dem Begriff der latenten Variable (oder auch *latent trait*) ist zumeist genau eine Variable gemeint. Mehrere latente Variablen werden auch als Konstrukte bezeichnet, da sie im Kontext einer Theoriebildung konstruiert worden sind (Rost, 2004, S. 30). Latente Variablen bzw. Konstrukte können indirekt über manifeste Indikatorvariablen messbar gemacht werden (Fischer, 1974, S. 148 ff.). Manifeste Variablen sind durch unmittelbare Beobachtung direkt messbar, wie beispielsweise Antworten auf Testaufgaben bzw. Testitems.[6] Latente Variablen können also über die Reaktion auf eine Aufgabe (beispielsweise richtige oder falsche Lösung) operationalisiert werden (vgl. Abbildung 3.4): „The mapping of observed performance onto the unobserved latent trait as well as the interpretation of individual students' observed performance in terms of the unobserved trait represent the types of activities that characterize latent trait analysis." (Ryan, 1983, S. 54)

Angewendet auf die Operationalisierung von Rechtschreibkompetenz bedeutet dies, dass die Schreibung von aufeinanderfolgenden Graphemen oder Graphemketten Hinweise auf die Beherrschung der Teilkompetenzen liefert. In Abbildung 3.5 ist ein Beispiel für die Messung des Prinzips der Wortbildung als Teilkompetenz des Rechtschreibtests SRT dargestellt. Die Schreibungen spezifischer Rechtschreibphänomene (zum Teil fett markiert) dienen hier als manifeste Indikatoren und stellen eine Auswahl der Analyseeinheiten dar. Dabei handelt es sich um die Wortbildungsmittel Derivation (durch die Schreibung der Affixe ⟨be⟩, ⟨Ver⟩, ⟨lich⟩ und ⟨chen⟩) und Komposition (Zusammen- bzw. Auseinanderschreibung von zwei Wortbestandteilen und Erkennen von Wortgrenzen bzw. aufeinanderfolgenden Morphemen (⟨ll⟩ und ⟨rr⟩)).

Die probabilistische Testtheorie wird als solche bezeichnet, da von Wahrscheinlichkeitszusammenhängen und -verteilungen ausgegangen wird. Es werden Abhängigkeiten zwischen den manifesten und latenten Variablen in den Modellannahmen formuliert. Über Itemcharakteristikfunktionen[7] können diese Abhängigkeiten, d. h. die Wahrscheinlichkeit einer richtigen Antwort oder Lösung einer Aufgabe in Abhängigkeit von der latenten Personen-

[6] Natürlich können damit auch Antworten auf Fragebogenitems gemeint sein, z. B. in Form von Zustimmung und Ablehnung von Aussagen. Da es in dieser Arbeit hauptsächlich um die Analyse von Tests und Auswertung von Testergebnissen geht, werden Itemreaktionen im Folgenden darauf bezogen.

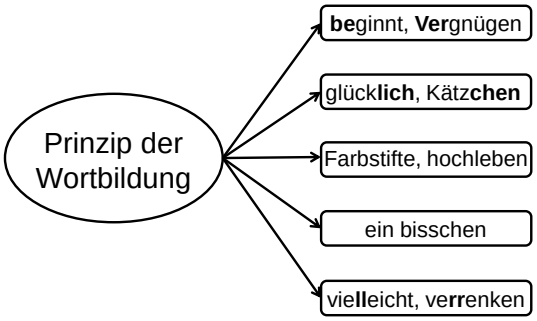



Abbildung 3.5: Beispiel für die Operationalisierung einer Teilkompetenz (in Anlehnung an Blatt, Voss, Kowalski & Jarsinski, 2011, S. 231)

fähigkeit, beschrieben werden (Carstensen, 2000, S. 18). Selbst bei Kenntnis der Ausprägungen einer latenten Variable einer Person kann nicht sicher die beobachtbare Reaktion vorhergesagt werden, sondern nur eine Wahrscheinlichkeit $p \in [0, 1]$ (Borg & Staufenbiel, 2007, S. 345). Personen mit hoher Personenfähigkeit werden demnach wahrscheinlich viele Items richtig beantworten (lösen, bejahen), wobei die Lösungswahrscheinlichkeit mit ansteigender Item- oder Aufgabenschwierigkeit sinkt. Aber auch Personen mit niedriger Personenfähigkeit haben bei schwierigen Items eine, wenn auch geringe, Lösungswahrscheinlichkeit (Gniewosz, 2011b, S. 75). Für die Analysen in Kapitel 4 bedeutet dies, dass Kinder mit einer hohen individuellen Merkmalsausprägung in Orthografie wahrscheinlich viele Wörter bzw. Wortbestandteile richtig schreiben werden, während leistungsschwache Schülerinnen und Schüler eine niedrigere Wahrscheinlichkeit für die normgerechte Schreibung der Analyseeinheiten besitzen.

### Grundlegende Unterscheidungsmerkmale von KTT und IRT

Die IRT unterscheidet sich von den traditionellen testtheoretischen Verfahren, die auf dem Ansatz der *klassischen Testtheorie* (KTT) basieren. Beide Ansätze sind allerdings weniger als konkurrierende, sondern vielmehr als sich ergänzende Modelle zu verstehen, wie u. a. Rost (2004, S. 12) aufzeigt. Molenaar (1995, S. 4) sieht die PTT als eine Weiterentwicklung der KTT mit größerem Potential an: „Briefly stated, IRT can do the same things better and can do more things, when it comes to modeling existing tests, constructing new ones, applying tests in non-standard settings, and above all interpreting the results of measurement."

Die KTT[8] ist der bisher am häufigsten verwendete Ansatz innerhalb der Psychometrie, deren Grundlagen bereits vor über einem Jahrhundert von Spearman (1904) entwickelt worden sind (Ingenkamp & Lissmann, 2008, S. 117). Aufbereitet wurde sie von Gulliksen (1950) und mathematisch beleuchtet von Lord und Novick (1968) (Moosbrugger, 2008b,

[7] Die Itemcharakteristikfunktion und -kurven für den Fall eines einparametrigen dichotomen Raschmodells werden weiter unten beschrieben.

[8] Nähere Ausführungen zur KTT sind z. B. in Lienert und Raatz (1998) nachzulesen.

S. 100; Fischer, 1974, S. 26). Zentraler Gegenstand der KTT ist die Frage nach dem Ausmaß von verfälschten Anteilen bei Messungen (Ingenkamp & Lissmann, 2008, S. 118). Daher wird die KTT auch Messfehlertheorie genannt. In ihrer Grundgleichung bestimmt sie den wahren Wert T (für true) für eine Person aus der Messwertvariable X und dem Messfehler E (für error) (de Gruijter & van der Kamp, 2008, S. 9 ff.; Rost, 2004, S. 36). Das Testergebnis entspricht damit direkt (unmittelbar) dem Personenmerkmal (wenn auch messfehlerbehaftet), d. h. die KTT postuliert einen deterministischen Zusammenhang zwischen dem Testergebnis und dem Personenmerkmal (Embretson & Reise, 2000, S. 42 f.). Die IRT geht, wie oben bereits erwähnt, nicht von direkt messbaren Merkmalen aus, sondern unterscheidet manifeste Variablen und latente Variablen, die in einem Wahrscheinlichkeitszusammenhang zueinander stehen (Ryan, 1983, S. 58). Sie betrachtet das Testergebnis lediglich als Indikator für ein latentes Merkmal, auf dessen Ausprägung es zu schließen gilt (Fischer, 1974, S. 148 ff.). „Das Wesen eines stochastischen Modells besteht darin, dass die Aussagen sich auf die Parameter von Zufallsverteilungen beziehen, nicht auf beobachtete Daten.“ (Fischer, 1968, S. 79)

Bei der KTT sind Testwerte stichprobenabhängig. So sind die Item- und Personenstatistiken von der untersuchten Stichprobe abhängig und können nicht mit anderen Stichproben verglichen und auf andere übertragen werden (Embretson & Reise, 2000, S. 25 ff.). Die Schwierigkeit von Aufgaben steigt in der KTT an, wenn eine Stichprobe von Personen mit niedrigen Fähigkeiten diese bearbeitet. Gleichzeitig sinkt die Anzahl der gelösten Aufgaben und damit der Fähigkeitswert der Personen. Analog dazu gilt, dass die Aufgabenschwierigkeiten sinken, wenn die Personen der Untersuchungsgruppe hohe Merkmalsausprägungen haben. In diesem Zuge werden viele Aufgaben richtig beantwortet und der Personenfähigkeitswert steigt. Der Schwierigkeitsindex in der KTT definiert sich aus dem prozentualen Anteil der Personen in einer Stichprobe, die die Aufgabe lösen (Lienert & Raatz, 1998, S. 32). In der PTT hingegen ermöglicht die Invarianzeigenschaft des Raschmodells (s. u.), Aussagen zu treffen, die von den Schwierigkeitsverteilungen der Items und den Personenfähigkeiten in anderen Stichproben unabhängig sind (Rost, 2004, S. 121).

Ein weiteres wichtiges Abgrenzungsmerkmal der PTT gegenüber der KTT ist die empirische Überprüfbarkeit der Gültigkeit eines Testmodells, die an dieser Stelle noch genannt werden soll, um den Wert der PTT für die empirische Bildungsforschung zu unterstreichen. Die PTT formuliert Modellannahmen über den Zusammenhang zwischen der beobachteten Reaktion auf Items und der Personenfähigkeit sowie Aufgabenschwierigkeit, deren Gültigkeit ermittelt werden kann (Moosbrugger, 2008a, S. 221; Borg & Staufenbiel, 2007, S. 315, 385; Bortz & Döring, 2006, S. 212; Carstensen, 2000, S. 20 ff.). Eine Testung über das Vorliegen des angenommenen Funktionsverlaufes, d. h. über die Passung zwischen Daten und Modell, ist damit möglich. Bei der KTT wird ein linearer Zusammenhang zwischen Item und latenter Eigenschaft postuliert, aber keine darauf basierende Funktion bestimmt, die empirisch überprüft werden kann (Carstensen, 2000, S. 20).

### 3.3.2 Das dichotome Modell von Rasch

Nun sollen die wesentlichen Eigenschaften und Annahmen des Raschmodells dargestellt werden, da dieses für die Auswertungen in Kapitel 4 die Basis bildet. Anhand der Ausführungen sollen die messtheoretischen Vorteile des Modells verdeutlicht werden. Das Raschmodell ist ein grundlegendes Testmodell der IRT und geht auf den Dänen Georg Rasch zurück, der es 1960 erstmals untersucht und dargestellt hat (Embretson & Reise, 2000, S. 48; Fischer, 1974, S. 199). Das Raschmodell gilt für dichotome Items[9] und beschreibt eine Wahrscheinlichkeitsfunktion aus Personenfähigkeit und Itemschwierigkeit (G. Rasch, 1960, S. 73 f.). Es trägt auch den Namen *einparametriges logistisches Modell* (1-pl, für one-parameter-logistic), da nur die Schwierigkeit als Parameter in das Modell eingeht[10] (Baker, 2001, S. 25): „Each model gives a probability distribution over a certain property of the event observed – number of misreadings; number of words read in a given time, or time used for reading the whole text; the answer to an item being correct or not – as determined by one parameter, which is, however, considered as the product of two factors, one referring to the test or item in question, the other referring to the person tested." (G. Rasch, 1960, S. 109)

#### Modellgleichung

Die von Rasch genannten zwei Faktoren haben Einfluss auf die Wahrscheinlichkeit der richtigen Beantwortung einer Aufgabe (Furr & Bacharach, 2008, S. 318). Ist eine Aufgabe leicht, wird sie wahrscheinlich von vielen Personen gelöst, da nur eine geringe Fähigkeit für die richtige Antwortreaktion benötigt wird. Ist eine Aufgabe schwer, erfordert die Lösung eine höhere Fähigkeitsausprägung, weshalb diese wahrscheinlich von weniger Personen korrekt beantwortet wird. Die Lösungswahrscheinlichkeit einer Aufgabe hängt also nur von der Differenz zwischen der Personenfähigkeit und der Aufgabenschwierigkeit ab (Fischer, 1974, S. 209). Dieser Zusammenhang ist in folgender Gleichung berücksichtigt:

$$p(X_{vi} = 1) = \frac{\exp(\theta_v - \sigma_i)}{1 + \exp(\theta_v - \sigma_i)} \tag{3.1}$$

(Rost, 2004, S. 119)

Gleichung 3.1 stellt die Modellgleichung des Raschmodells dar. Sie drückt die Lösungswahrscheinlichkeit $p$ für eine richtige Antwort ($X_{vi} = 1$) einer Person $v$ (mit Personenfähigkeit $\theta_v$) des Items $i$ (mit Aufgabenschwierigkeit $\sigma_i$) aus.[11] Da es sich um eine Wahrscheinlichkeitsfunktion handelt, nimmt die Gleichung Werte zwischen 0 (sicher falsche

[9] Dichotome oder binäre Items sind zweistufige Antworten und lauten bei Fragebogenitems z. B. „ja" und „nein" oder bei Testitems „richtig" und „falsch" oder „gelöst" und „nicht gelöst" (Furr & Bacharach, 2008, S. 46 ff.).

[10] Eine Beschreibung des zwei- und dreiparametrigen-Modells findet sich z. B. bei Embretson und Reise (2000, S. 70 ff.).

[11] $\exp(x)$ bezeichnet hier die Exponentialfunktion mit der eulerschen Zahl $e \sim 2{,}72$ als Basis.

Beantwortung des Items) und 1 (sichere Lösung des Items) an (Ryan, 1983, S. 57; Wright & Stone, 1979, S. 15, 17). Die Funktion wurde abgeleitet aus:

$$\log \frac{p(X_{vi} = 1)}{p(X_{vi} = 0)} = \theta_v - \sigma_i \tag{3.2}$$

(Rost, 2004, S. 118)

In Gleichung 3.2 ist die Lösungswahrscheinlichkeit in logit-transformierter Form dargestellt. Ein Logit ist der Logarithmus[12] des Quotienten aus Lösungswahrscheinlichkeit und Gegenwahrscheinlichkeit (Rauch & Hartig, 2008, S. 241). Hier zeigt sich, dass die Logits der Lösungswahrscheinlichkeiten eine lineare Funktion der Personenfähigkeit und der Aufgabenschwierigkeit sind (Rost, 2004, S. 118). Da der Aufgabenparameter vom Personenparameter abgezogen wird, drückt $\sigma_i$ die Aufgabenschwierigkeit und nicht -leichtigkeit aus (Walter, 2005, S. 33). Ist eine Aufgabe schwer, ist der Logit der Lösungswahrscheinlichkeit gering.

Die Wahrscheinlichkeit für die richtige Beantwortung einer Aufgabe, in Abhängigkeit von der individuellen Merkmalsausprägung in einer latenten Variable und der Schwierigkeit, lässt sich durch den Einsatz von Zahlen in Gleichung 3.1 konkret veranschaulichen. So beträgt diese z. B. für eine (fähige) Person $v$ mit $\theta_v = 2$ und einem (leichten) Item $i$ mit $\sigma_i = -0,5$:

$$p(X_{vi} = 1 \mid \theta_v = 2, \sigma_i = -0,5) = \frac{\exp(2 - (-0,5))}{1 + \exp(2 - (-0,5))} = 0,92$$

Damit hat die Person eine (hohe) Lösungswahrscheinlichkeit für das Item von 92 Prozent.

### Itemcharakteristikkurve

Der Funktionsverlauf des Raschmodells (Gleichung 3.1) kann über eine *Itemcharakteristikkurve* (kurz ICC, vom englischen Item Characteristic Curve) grafisch visualisiert werden. Abbildung 3.6 zeigt die Beziehung der Lösungswahrscheinlichkeit einer Aufgabe in Abhängigkeit von der Ausprägung der Personenfähigkeit an (Embretson & Reise, 2000, S. 46 f.). Da es sich in Gleichung 3.1 um einen logistischen Zusammenhang handelt, ist der Funktionsverlauf s-förmig und nähert sich asymptotisch den Schranken 0 und 1 an (Geiser & Eid, 2010, S. 313). Auf der Abszisse sind die Werte der Personenfähigkeit $\theta$ und auf der Ordinate die Lösungswahrscheinlichkeiten für eine richtige Antwort $p(X_{vi} = 1)$ abgetragen. Anhand der Itemfunktion lässt sich ablesen, dass eine Person mit einer höheren Fähigkeit, beispielsweise von $\theta_v = 6$, eine größere Wahrscheinlichkeit hat, die Aufgabe zu lösen, als eine Person mit einer niedrigeren Fähigkeit von z. B. $\theta_v = 2$. Sind Personenfähigkeit und Aufgabenschwierigkeit gleich groß, beträgt die Antwortwahrscheinlichkeit 50 Prozent (Bortz & Döring, 2006, S. 208). Generell gilt: Ist eine Person fähiger als eine Aufgabe

[12] Parallel zu Rost (2004, S. 118 f.) ist mit $\log(x)$ der Logarithmus zur Basis $e$ gemeint (natürlicher Logarithmus, logarithmus naturalis, ln).

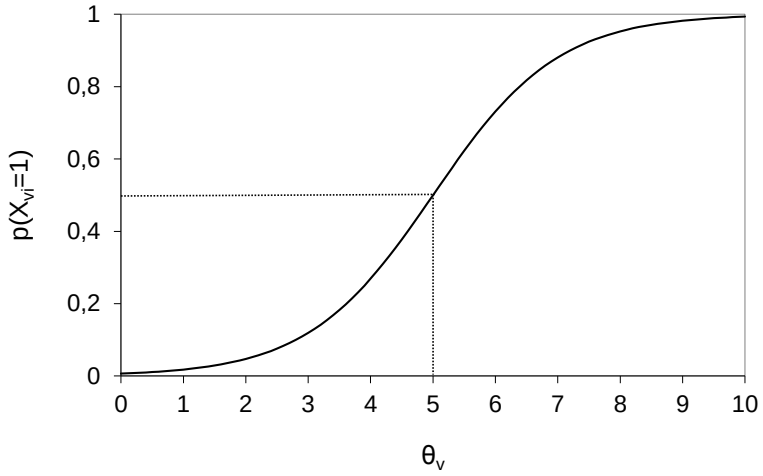



Abbildung 3.6: ICC für eine Aufgabe $i$ mit Schwierigkeit $\sigma_i = 5$ (in Anlehnung an Bond & Fox, 2007, S. 46)

schwer ist, also $\theta_v > \sigma_i$, steigt die Lösungswahrscheinlichkeit. Die Gegenrichtung gilt analog: Ist eine Aufgabe schwerer als eine Person fähig ist, also $\theta_v < \sigma_i$, sinkt die Wahrscheinlichkeit (G. Rasch, 1960, S. 73). Wie Abbildung 3.6 zeigt, werden Itemschwierigkeit und Personenfähigkeit bei der PTT auf einer gemeinsamen Skala abgebildet (Wu & Adams, 2007, S. 62), die bei IRT-Modellen mit logistischer ICC auch Logit-Skala genannt wird (Rauch & Hartig, 2008, S. 241). So kann die Schwierigkeit eines Items direkt bei der Fähigkeitsausprägung abgelesen werden, für die die Lösungswahrscheinlichkeit 50 Prozent beträgt (vgl. die gestrichelten Linien in Abbildung 3.6) (Geiser & Eid, 2010, S. 313; Furr & Bacharach, 2008, S. 316). Zudem sind direkte Vergleiche zwischen Personen, zwischen Items sowie zwischen Personen und Items möglich, um Differenzen (in Leistung oder Schwierigkeit) auszumachen (Bond & Fox, 2007, S. 47).

Da ein Test nicht nur aus einer Aufgabe besteht, finden sich in Abbildung 3.7 mehrere ICCs. Hier tritt eine Besonderheit des Raschmodells zum Vorschein: Die Itemfunktionen sind verschoben, da sie unterschiedliche Schwierigkeitsgrade aufweisen ($\sigma_1 = 4$, $\sigma_2 = 5$, $\sigma_3 = 6$), sie verlaufen aber parallel und überschneiden sich demnach nie (Fischer, 1974, S. 199). Je weiter links die ICC liegt, desto leichter ist die Aufgabe bzw. je weiter rechts sie liegt, desto schwieriger, wobei die Form der Kurve stabil bleibt (Furr & Bacharach, 2008, S. 325). Diese Eigenschaft ergibt sich aus Gleichung 3.1 des Raschmodells. Für jede Aufgabe gibt es zwar einen Schwierigkeitsparameter, der variabel ist (der also bestimmt, wie weit die Funktion links oder rechts liegt), es existiert aber kein Parameter, der die Steigung oder Form verändert (Strobl, 2010, S. 11). Dies bedeutet gleichzeitig, dass allen Aufgaben gemein ist, dass sie die gleiche Trennschärfe besitzen (Rost, 2004, S. 120).

Abbildung 3.8 zeigt zwei Aufgaben mit unterschiedlicher Trennschärfe. Wird der Trennschärfeparameter $\beta_i$, der die Steigung des Funktionsverlaufs bestimmt, in die Gleichung des Raschmodells integriert, wird vom zweiparametrigen logistischen Modell (2-pl) gesprochen (de Gruijter & van der Kamp, 2008, S. 136). Die Itemfunktion mit durchgezogener

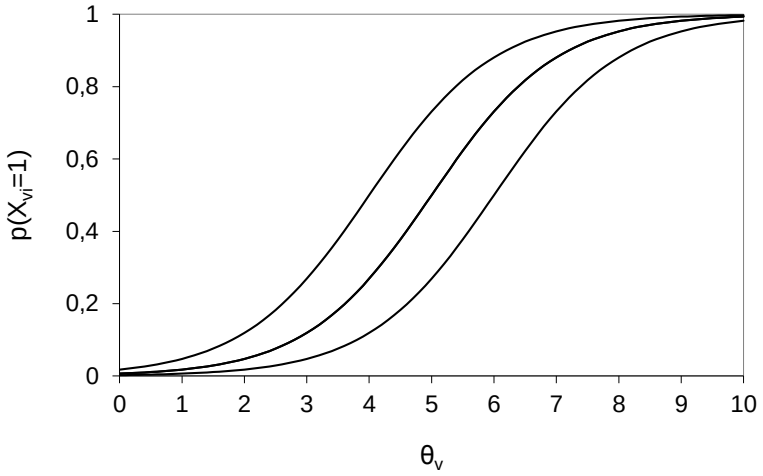



Abbildung 3.7: Parallele ICCs für drei Aufgaben mit unterschiedlichen Schwierigkeiten (in Anlehnung an Bond & Fox, 2007, S. 63)

Linie (Item 1) hat eine Trennschärfe von $\beta_1 = 0,5$, während Funktion 2 (gestrichelte Linie) eine höhere Trennschärfe von $\beta_2 = 2$ aufweist (der Schwierigkeitsparameter beträgt jeweils 5). Anhand von Abbildung 3.8 lassen sich folgende Punkte verdeutlichen:

- Aufgaben mit hoher Trennschärfe (Itemfunktion 2) weisen für Personen mit leicht unterschiedlichen Personenfähigkeiten (die den Aufgabenschwierigkeiten angemessenen sind) deutlichere Unterschiede in den Lösungswahrscheinlichkeiten auf als Aufgaben mit niedriger Trennschärfe (Funktion 1) (Embretson & Reise, 2000, S. 47). Bei Item 1 beträgt die Differenz der Lösungswahrscheinlichkeiten zwischen Person *b* (mit $\theta_b = 5,5$) und Person *c* (mit $\theta_c = 6$) ca. 5 Prozentpunkte, während sie bei Item 2 mit rund 15 Prozentpunkten höher ausfällt. Je steiler eine Itemfunktion ansteigt, desto stärker (trennschärfer) können geringe Unterschiede in der Personenfähigkeit stark unterschiedlichen Lösungswahrscheinlichkeiten zugeordnet werden (Moosbrugger, 2008a, S. 238; Rost, 2004, S. 98).
- Das trennschärfere Item 2 liefert wenige bis keine Informationen über Personen in den Randbereichen der Fähigkeiten. Auf dieses Problem wird im Rahmen der Modelloptimierung in Abschnitt 3.3.2.2 eingegangen.
- Die Reihenfolge der Lösungswahrscheinlichkeiten kann sich bei Items mit unterschiedlichen Trennschärfen abwechseln. Personen mit Fähigkeiten < 5 haben eine höhere Wahrscheinlichkeit, Item 1 richtig zu beantworten als Item 2 richtig zu beantworten. Demgegenüber weisen Personen mit Fähigkeiten > 5 eine höhere Wahrscheinlichkeit für Item 2 als für Item 1 auf. Das bedeutet, dass Itemschwierigkeiten im 2-pl Modell von der Stichprobe abhängig sind (Rost, 2004, S. 134).

Da die Gleichung des Raschmodells (Gleichung 3.1) ausschließlich die Parameter der Personenfähigkeit und Aufgabenschwierigkeit beinhaltet, stellen Aufgaben mit unterschiedlichen Trennschärfen eine Verletzung der Modelleigenschaften dar. Das im Folgenden beschriebene Kriterium der spezifischen Objektivität wäre dann nicht mehr erfüllt.

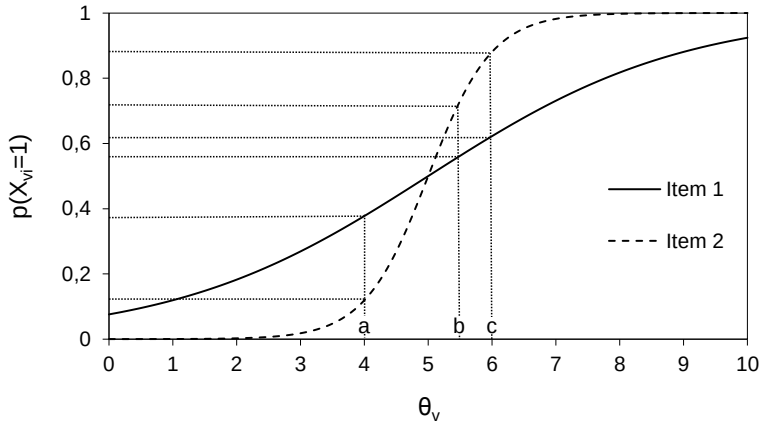



Abbildung 3.8: ICCs für zwei Aufgaben mit unterschiedlichen Trennschärfen (in Anlehnung an de Gruijter & van der Kamp, 2008, S. 137; Rost, 2004, S. 134)

### Invarianzeigenschaft – Spezifische Objektivität

Die Invarianz oder spezifische Objektivität ist eine elementare Eigenschaft des Raschmodells (Fischer, 1974, S. 407 ff.). Obwohl Personenfähigkeit und Aufgabenschwierigkeit auf einer gemeinsamen Skala ausgewiesen werden (vgl. Abbildung 3.6), werden sie getrennt voneinander als eigenständige Größen aus den Testdaten geschätzt. Das heißt, die Invarianzeigenschaft beschreibt die Unabhängigkeit der Personenparameter von den ausgewählten Items aus einem Item-Universum sowie die Unabhängigkeit der Itemparameter von der Stichprobe der Personen (Rost, 2004, S. 40, 121 f.). Möglich ist dies durch die Annahme, dass die Form der ICCs immer gleich ist und sie lediglich parallel an der Abszisse verschoben sind (vgl. Abbildung 3.7). Die Invarianzeigenschaft soll im Folgenden kurz ausgeführt und anhand von Beispielen verdeutlicht werden.

Da die Items homogen sind (Homogenitätseigenschaft, s. u.), ist es egal, welche Aufgaben für die Schätzung der Personenfähigkeiten genutzt werden: „Jede Teilmenge an Items führt also im Allgemeinen zu den gleichen Personenscores." (Borg & Staufenbiel, 2007, S. 350) Schwerere und leichtere Aufgaben verändern, im Gegensatz zur KTT (s. o.), nicht die Fähigkeitswerte der Stichprobe. Die Personeneigenschaften sind also unveränderlich gegenüber den Items. Zudem sind die Personenfähigkeiten von der Aufgabe, über die der Vergleich stattfindet, unabhängig. Bei Aufgaben mit unterschiedlichen Schwierigkeiten – die aber ansonsten homogen (eindimensional) bezüglich der Personenfähigkeit sind – hat eine fähigere Person stets eine höhere Lösungswahrscheinlichkeit als eine weniger fähige Person. Diesen spezifisch objektiven Vergleich verdeutlicht Abbildung 3.9. Die Lösungswahrscheinlichkeit von Person $b$ (mit $\theta_b = 7$) ist immer größer als von Person $a$ (mit $\theta_a = 4$). Würden weitere parallele Funktionsverläufe in Abbildung 3.9 eingezeichnet, bliebe diese Ordnung trotzdem erhalten (Strobl, 2010, S. 20).

Für die Itemparameter im Raschmodell gilt analog das Gleiche. Die Aufgabenschwierigkeiten sind unveränderlich gegenüber den Fähigkeitsausprägungen in der Stichprobe (Fischer,

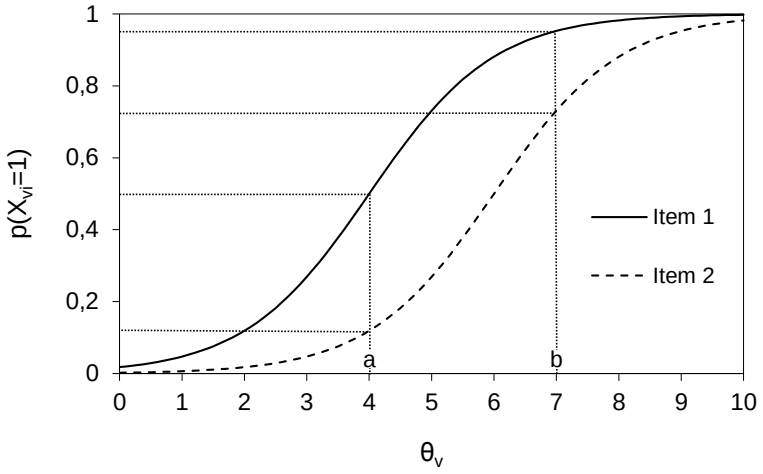



Abbildung 3.9: Spezifische Objektivität beim Vergleich der Personenfähigkeiten und Itemschwierigkeiten (in Anlehnung an Borg & Staufenbiel, 2007, S. 349)

1974, S. 228; Fischer, 1968, S. 72). Während bei der KTT eine Stichprobe mit besonders (un)fähigen Personen die Aufgabenschwierigkeit beeinflusst (senkt oder steigert; s. o.), bleibt diese im Raschmodell konstant: „the best estimate of the degree of difficulty of an item is found by using its frequency of correct answers in the whole population, irrespective of which persons have solved the item correctly.“ (G. Rasch, 1960, S. 77) Zudem sind die Aufgabenschwierigkeiten von den Personen, über die der Vergleich stattfindet, unabhängig. Bei verschiedenen Personen mit niedrigen oder hohen Merkmalsausprägungen hat ein schweres Item immer eine niedrigere Lösungswahrscheinlichkeit als ein leichtes. Aufgabe 1 in Abbildung 3.9 ist für die Personen *a* und *b* am einfachsten, während die Lösungswahrscheinlichkeit für Aufgabe 2 für beide Personen geringer ist. Im Gegensatz dazu hat in dem 2-pl-Modell in Abbildung 3.8 die Stichprobe Einfluss auf den Vergleich zweier Aufgaben: so ist bei Person *a* die Lösungswahrscheinlichkeit von Item 1 höher als von Item 2, während Person *c* eine höhere Lösungswahrscheinlichkeit bei Item 2 als bei Item 1 aufweist.

### Eindimensionalität

Die Eindimensionalität der Items ist eine weitere Annahme des Raschmodells. Sie bedeutet, dass alle Items dieselbe Personenfähigkeit erfassen (Bond & Fox, 2007, S. 134; Ryan, 1983, S. 50). Ist dies der Fall, so werden diese Items als homogen bezeichnet. Beispielsweise sollten die Items in einem Test der Mathematikkompetenz nur Mathematikkompetenz erheben, nicht auch noch z. B. Sprachkompetenz (Strobl, 2010, S. 23). Aus der Homogenitätseigenschaft der Items ergibt sich, dass die Anzahl der gelösten Aufgaben eine suffiziente Statistik für die Personenfähigkeit darstellt. Mit einem Test werden im Rahmen von mehrdimensionalen Raschmodellen auch Konstrukte mit mehreren Dimensionen erhoben (s. u.).

### Suffiziente Statistiken

Im Raschmodell werden für jeden Parameter (sowohl Aufgabenschwierigkeits- als auch Personenfähigkeitsparameter) suffiziente oder erschöpfende Statistiken ausgegeben (Strobl, 2010, S. 15; Molenaar, 1995, S. 10 f.). Man stelle sich eine Datenmatrix vor, in der in den Zeilen die Personen und in den Spalten die Items abgetragen sind. Für jede Person enthält die Zeilenrandsumme vollständig alle Informationen über den Personenparameter. Entsprechend umfasst die Spaltenrandsumme für jede Aufgabe die gesamte Information über den Aufgabenparameter. Da die Zeilenrandsummen suffiziente Statistiken darstellen, ist es nicht wichtig, welche (schwierigen oder leichten) Aufgaben von einer Person gelöst worden sind, sondern nur wie viele (Rost, 2001, S. 27; Embretson & Reise, 2000, S. 57; G. Rasch, 1960, S. 76 f.). So werden beispielsweise Personen mit den beiden unterschiedlichen Antwortmustern 0010111101 und 1110101001 nicht unterschieden. Die Anzahl der richtig beantworteten Items (in dem Beispiel 6) enthält alle benötigten Informationen für die Schätzung der Personenparameter (Fischer, 1968, S. 101). Alle Personen mit identischer Summe gelöster Aufgaben (Itemscore) erhalten denselben Personenparameter (Rost, 2004, S. 113).

Für eine einzelne Aufgabe ergibt dies natürlich keinen Sinn. Wird aber ein Pool von Items betrachtet, so werden Personen mit niedrigerer Fähigkeit hauptsächlich die leichteren Items lösen und Personen mit höherer Fähigkeit diese leichteren sowie zusätzlich schwierigere Items. Somit fällt die Anzahl der insgesamt gelösten Aufgaben bei Personen mit hoher Fähigkeitsausprägung größer aus. Natürlich kann es sein, dass Personen mit niedriger Fähigkeit, z. B. zufallsbedingt, auch schwere Aufgaben richtig beantworten. Vom Raschmodell wird entsprechend eine Lösungswahrscheinlichkeit vorausgesagt, die aber im Vergleich zu Personen mit hoher Merkmalsausprägung eben klein bzw. geringer ausfällt (Strobl, 2010, S. 16).

### Lokale stochastische Unabhängigkeit

Beim Raschmodell wird für eine Person (bzw. für mehrere Personen mit gleicher Fähigkeitsausprägung) die Unabhängigkeit der Lösungswahrscheinlichkeit eines Items von allen anderen Items angenommen. Diese Annahme wird als lokale stochastische Unabhängigkeit[13] bezeichnet (Embretson & Reise, 2000, S. 48). Sie basiert darauf, dass sich die gemeinsame Wahrscheinlichkeit von stochastisch unabhängigen Ereignissen als Produkt der Einzelwahrscheinlichkeiten darstellen lässt (Bühner & Ziegler, 2009, S. 125).

Die Wahrscheinlichkeit, dass eine Person eine Aufgabe richtig oder falsch beantwortet, hängt ausschließlich von der Schwierigkeit und Fähigkeit ab (daher auch die Bezeichnung „lokal“) (Fischer, 1974, S. 211). Daraus lässt sich ableiten, dass sich die Wahrscheinlichkeit einer Person, alle Aufgaben zu lösen, aus dem Produkt der Wahrscheinlichkeiten für

[13] Es handelt sich hierbei um keine alleinige Eigenschaft des Raschmodells, aber um ein übliches Merkmal von Latent-Trait- und Latent-Class-Modellen (Rost, 2001, S. 29).

die Lösung der einzelnen Aufgaben ergibt. Die Lösungswahrscheinlichkeit darf sich in dem Modell nicht durch das Lösen anderer Items verändern. Dies trifft z. B. bei Tests zu, deren Aufgaben inhaltlich aufeinander aufbauen, wenn also die Lösung von Frage A Voraussetzung für die Beantwortung von Frage B darstellt. Hier würde die Lösungswahrscheinlichkeit von Frage B automatisch auf 0 sinken, wenn keine Antwort auf Frage A vorliegt (Strobl, 2010, S. 18).

### Mehrdimensionale Raschmodelle

Beim mehrdimensionalen Raschmodell wird das Raschmodell bezüglich der Anzahl an latenten Dimensionen erweitert (Walter & Rost, 2011, S. 116). Dies bedeutet, dass mehrere latente Personenfähigkeiten zusammenkommen, um die Items eines Tests richtig zu lösen. Das eindimensionale Raschmodell gilt innerhalb jeder Dimension des mehrdimensionalen Modells, um das Antwortverhalten zu beschreiben (Bond & Fox, 2007, S. 259). Viele Tests nutzen latente Variablen, die nicht nur eine latente Dimension abbilden, sondern mehrere. Beispielsweise wurden in der Konzeption des Naturwissenschaftstests in PISA sieben Teilkompetenzen voneinander unterschieden. Einige Kinder lösen vornehmlich die Aufgaben der Teilkompetenz *Umgang mit Graphiken*, während andere sicherer in *Mentale Modelle* (hier geht es um die Nutzung räumlich-geometrischer Vorstellungen) sind und in dieser Teilkompetenz viele Aufgaben richtig beantworten (Rost, Walter, Carstensen, Senkbeil & Prenzel, 2004, S. 122 ff.). Daher wird naturwissenschaftliche Kompetenz nicht über eine globale Dimension, sondern über mehrere Dimensionen abgebildet. Entsprechend wird Rechtschreibkompetenz in gutschrift und im SRT über die Leistungen in mehreren Teilkompetenzen erhoben, und nicht ausschließlich über das Auszählen vollständig richtig geschriebener Wörter.

Die Anzahl der latenten Dimensionen und die Zuordnung der Items zu den Dimensionen wurde für die beiden Rechtschreibtests bereits im Voraus festgelegt (konfirmatorisches Verfahren). Die Bestimmung erfolgte auf Basis der jeweiligen theoretischen Annahmen über die Struktur von Rechtschreibkompetenz. Mehrdimensionale Modelle können Antworten auf Fragen zur Struktur von Kompetenz geben, indem sie die Dimensionalität der Daten überprüfen und das beste Modell, unter mehreren konkurrierenden Alternativmodellen, mit Hilfe unterschiedlicher Statistiken (die in Abschnitt 3.3.2.1 beschrieben werden), bestimmen (Wu & Adams, 2006, S. 104; Carstensen, 2000, S. 23). Einen wichtigen Bestandteil im Modellprüfungsprozess bilden dabei die latenten Korrelationen der einzelnen Dimensionen bzw. Teilkompetenzen untereinander, die häufig hoch ausfallen, da die Personen zumeist in einer Teilkompetenz als auch den restlichen eher besser oder eher schlechter abschneiden (Hartig & Höhler, 2010, S. 193). Gleichzeitig bieten mehrdimensionale Modelle differenzierte Hinweise zur Förderung von Schülerinnen und Schülern (Hartig & Höhler, 2010, S. 190).

Im Rahmen von mehrdimensionalen Modellen werden *between-item multidimensionality* und *within-item multidimensionality* unterschieden. Bei between-item multidimensionality wird jede Aufgabe genau einer latenten Dimension zugeordnet. Dies bedeutet, dass jedes

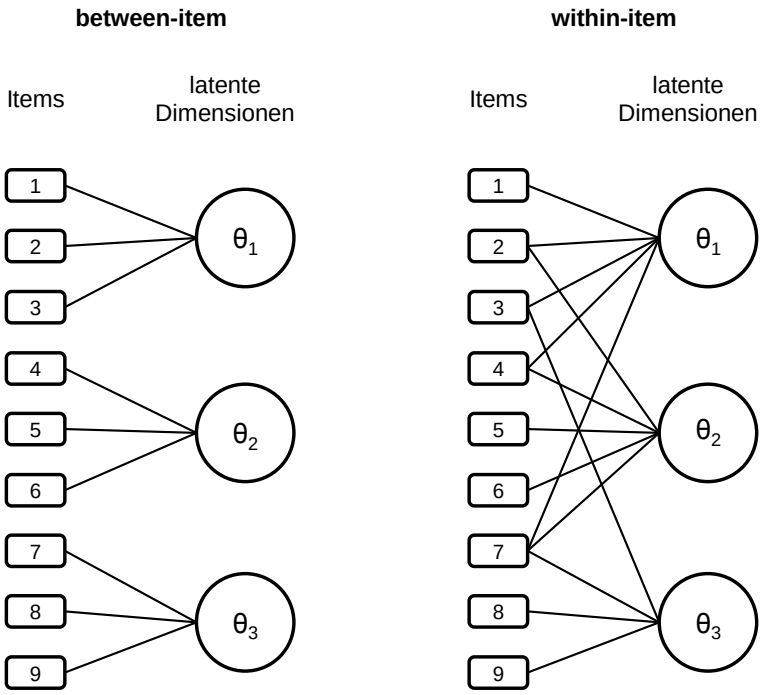



Abbildung 3.10: Between-item und within-item multidimensionality (in Anlehnung an Adams, Wilson & Wang, 1997, S. 9)

Item genau eine latente Variable misst und damit disjunkte eindimensionale Subskalen vorhanden sind: „In such tests, each item belongs to only one particular subscale, and there are no items in common across the subscales.“ (Adams, Wilson & Wang, 1997, S. 11) Bei within-item multidimensionality werden Aufgaben mehreren latenten Dimensionen zugeordnet, sodass mit einem Item gleichzeitig mehrere Personenmerkmale erhoben werden (Walter, 2005, S. 37).

Der Unterschied zwischen den beiden Fällen von Mehrdimensionalität ist in Abbildung 3.10 veranschaulicht. Bei dem between-item-multidimensional-Modell auf der linken Seite laden drei der neun Items jeweils und ausschließlich auf einer der drei latenten Dimensionen. Die latenten (Sub-)Dimensionen repräsentieren separate Fähigkeiten oder Teilkompetenzen. Die mehrdimensionale Modellierung von Rechtschreibkompetenz erfolgt in Kapitel 4 über between-item multidimensionality, da die Kompetenzen jeweils separate Dimensionen erfassen und dementsprechend jede Analyseeinheit in den beiden Tests jeweils und ausschließlich einer Teilkompetenz zugeordnet ist (vgl. Abschnitte 2.2 und 2.3). Das within-item-multidimensional-Modell, das sich auf der rechten Seite von Abbildung 3.10 befindet, zeigt eine komplexere Ladungsstruktur. Die Beantwortung eines Items wird hier von mehreren latenten Dimensionen beeinflusst (Adams & Wu, 2007, S. 62). Item 7 dient beispielsweise als Indikator für alle drei Dimensionen.

Analog zu dem Vorgehen in den großen Schulleistungsstudien wird das *multidimensional random coefficients multinomial logit Modell* (MRCML-Modell) für die mehrdimensio-

nalen Analysen in Kapitel 4 verwendet.[14] Es handelt sich um ein Modell, welches u. a. das 1-pl-Modell als auch das *Partial-Credit-Modell* (vgl. Abschnitt 4.2.1) integriert. Mit ihm können sowohl between-item multidimensionality als auch within-item multidimensionality modelliert werden (Adams et al., 1997, S. 9 ff.). Das MRCML-Modell ist in der Software „ConQuest“ implementiert, die in Kapitel 4 für die Auswertung der Leistungsdaten verwendet wird (Wu, Adams, Wilson & Haldane, 2007, S. 133 ff.).

### Parameterschätzung

Wie am Anfang von Abschnitt 3.3.2 erläutert, stellen Personenfähigkeiten und Aufgabenschwierigkeiten die zwei zu ermittelnden Größen beim Raschmodell dar (vgl. Modellgleichung 3.1). Da es sich um mehrere unbekannte Modellparameter handelt, ist es nicht möglich, eine Gleichung zu berechnen bzw. aufzulösen, weshalb Schätzungsverfahren angewendet werden müssen (Rost, 2004, S. 301). Die *Maximum-Likelihood-Methode*[15] ist ein einschlägiges Verfahren zur Parameterschätzung, das im Rahmen der Datenauswertung in Kapitel 4 verwendet wird.

Grundlage für die Parameterschätzung ist die *Likelihoodfunktion*: „Die Likelihoodfunktion beschreibt die Wahrscheinlichkeit der Daten unter der Annahme, dass das gewählte Modell gilt.“ (Rost, 2004, S. 303) Damit hat die Likelihoodfunktion zwei Einsatzgebiete: zum einen kann sie für die Parameterschätzung genutzt und zum anderen für die Modellgeltungsprüfung (goodness-of-fit-Test) herangezogen werden. Die Likelihoodfunktion setzt sich aus den gesamten Testdaten zusammen, indem sie die Wahrscheinlichkeiten der einzelnen Itemantworten über alle Personen (Zeilen) und Items (Spalten) aufmultipliziert (Rost, 2004, S. 112 f.). Voraussetzung hierfür ist die lokale stochastische Unabhängigkeit (s. o.) der Testergebnisse (Borg & Staufenbiel, 2007, S. 359 f.). Als Konvention hat sich etabliert, die Likelihoodfunktion zu logarithmieren[16] (log-Likelihood), um das Schätzverfahren zu vereinfachen (Gautschi, 2010, S. 210; Embretson & Reise, 2000, S. 162). Dadurch bleibt der Ort des Maximums gleich, aber die log-Likelihood nimmt Werte zwischen $-\infty$ und 0 an (Rost, 2004, S. 304 f.). Je weiter sich der Likelihoodwert 0 annähert, umso besser ist er, d. h. umso wahrscheinlicher ist das Modell für die Daten gültig. Ziel und Prozedur der Parameterschätzung mit der Maximum-Likelihood-Methode ist diese Annäherung. Sie erfolgt durch ein iteratives Verfahren, in dem Rechenschritte wiederholt werden. Geschätzte Näherungswerte der unbekannten Modellparameter werden dabei als neue Anfangswerte für die aktuelle Schätzung genutzt (Wu & Adams, 2007, S. 63; Baker, 2001, S. 48). Das Verfahren wird wiederholt, bis ein definiertes Abbruchkriterium greift bzw. ein bestimmtes Optimierungskriterium erfüllt ist (wenn sich beispielsweise nur noch kleinste Veränderungen der log-Likelihood ergeben), da ansonsten unendlich lang weitergerechnet würde (Gollwitzer, 2008, S. 289).

[14] Für eine ausführliche Beschreibung sowie Modellspezifikationen sei hiermit auf Adams et al. (1997) verwiesen.

[15] Weitere Details zur Maximum-Likelihood-Methode finden sich u. a. in Gautschi (2010), de Gruijter und van der Kamp (2008, S. 149), Walter (2005, S. 41 ff.), Embretson und Reise (2000, S. 200 ff.).

[16] Auch hier wird üblicherweise der natürliche Logarithmus verwendet.

Eine derart optimierte Likelihood wird als maximale Likelihood bezeichnet. Die unbekannten Modellparameter werden also so lange geschätzt, bis die Schätzungen konvergieren, also zwischen den Iterationen keine bedeutenden Veränderungen mehr erzeugt werden (Embretson & Reise, 2000, S. 206). Bei der Maximum-Likelihood-Methode werden damit alle Modellparameter auf den Wert festgelegt, an dem die Likelihoodfunktion ihr Maximum[17] hat (Rost, 2004, S. 304). Je größer die Likelihood ausfällt, umso höher ist die Wahrscheinlichkeit, mit den geschätzten Modellparametern genau die Daten zu bekommen, die beobachtet wurden (Gautschi, 2010, S. 207; Gollwitzer, 2008, S. 289).

#### 3.3.2.1 Modellgeltungstests

Mit Modellgeltungstests wird die Frage nach der Modellkonformität beantwortet, also inwieweit sich die theoretischen Modelle auf die beobachteten Daten anwenden lassen. Rein theoretisch lässt sich jedes IRT-Modell auf Daten anwenden, da die Parameter auch dann geschätzt werden, wenn sie schlecht auf die Daten passen (Burnham & Anderson, 2004, S. 262). Daher ist die interessante Frage nicht die danach, ob die Daten auf das Modell passen, sondern wie gut die Daten passen (Rost, 2004, S. 330). Oder anders ausgedrückt: wie gering die Abweichungen sind und ob diese zufällig zustande kommen. Es kann also eine Aussage darüber getroffen werden, ob Modell A besser die Daten erklärt als Modell B. Es existiert nicht das eine wahre Modell, da Modelle immer (restriktive) mathematisch-abstrakte Abbildungen der (komplexen) Realität sind: „there are no procedures that result in a researcher stating *definitively* that a particular model does or does not fit, or is or is not appropriate.“ (Embretson & Reise, 2000, S. 233)

Die zentrale Rolle für die Beurteilung der Güte der Datenanpassung spielen das *Einfachheitskriterium* und das *Geltungskriterium*. Das Einfachheitskriterium besagt, dass eine Theorie umso besser ist, je einfacher sie ist (Rost, 2004, S. 330). Gleichzeitig muss der Geltungsbereich der Theorie stimmig sein. Eine Theorie mit wenigen und einfachen Annahmen muss genauso valide sein wie eine komplexere Theorie: sie muss dieselben Sachverhalte beschreiben und erklären. Der Geltungsbereich der Theorie ist bei den Analysen in Kapitel 4 gegeben, da sich alle berechneten Modelle auf denselben Datensatz beziehen. Rost (2004, S. 331) fasst zusammen: „Die Prüfung der Modellgültigkeit hat daher drei Dinge im Auge zu behalten, nämlich erstens, wie gut erklärt das Modell die Daten, zweitens, mit welchem Aufwand an Modellparametern wird dies erreicht und drittens, wie gut passt das Modell zum Forschungsstand in dem Gebiet.“ Neben statistischen spielen damit gleichzeitig auch inhaltliche Überlegungen des spezifischen Fachgebiets eine Rolle.

[17] Bei Raschmodellen wird von Likelihoodfunktionen ohne lokale Maxima ausgegangen, während diese bei latenten Klassenanalysen bedacht werden müssen (Rost, 2004, S. 307). Im Rahmen der probabilistischen Clusteranalyse (vgl. Abschnitt 3.3.3) wird daher noch einmal gesondert auf sie eingegangen.

### Likelihoodquotiententest

Der Likelihoodquotiententest (likelihood ratio test, LR) für das Raschmodell ist im Vergleich zu den informationstheoretischen Maßen (s. u.) ein inferenzstatistischer Modelltest. Mit ihm kann eine Aussage über die globale Geltung des Modells getroffen werden (Fischer, 1974, S. 300). Diese Aussage ist statistisch abgesichert. Dadurch kann eindeutig bestimmt werden, welches von zwei alternativen Modellen besser auf die Daten passt (Rost, 2004, S. 339). Für die Durchführung des LR-Tests werden die Likelihoods zweier konkurrierender Modelle verglichen, indem der Quotient aus beiden gebildet wird. Der Likelihoodquotient bildet sich also aus:

$$LR = \frac{L_0}{L_1} \tag{3.3}$$

(Rost, 2004, S. 332)

Wichtig ist dabei, dass sich die Likelihoodwerte auf dieselbe Datenmatrix beziehen müssen (Rost, 2004, S. 330). Die Daten müssen identisch sein, es darf z. B. kein Item und keine Person hinzukommen oder wegfallen. Zudem muss die Voraussetzung erfüllt sein, dass die Likelihood, die sich im Nenner befindet ($L_1$), ein Obermodell von dem Modell im Zähler ist (damit ist $L_0$ ein Untermodell) (Carstensen, 2000, S. 169). Das Untermodell ergibt sich also aus einer Restriktion der Parameter des Obermodells (Borg & Staufenbiel, 2007, S. 363).[18] Bei dem Vergleich verschiedener alternativer Rechtschreibkompetenzmodelle in Abschnitt 4.2 geht beispielsweise das eindimensionale Modell aus den mehrdimensionalen Modellen (allgemeinere Modelle) hervor und bildet daher das restriktivere Modell (das im Zähler steht).

Der Likelihoodquotient lässt sich in eine $\chi^2$-verteile Prüfgröße durch Logarithmieren und Multiplikation mit dem Faktor $-2$ überführen (Fischer, 1974, S. 369):

$$-2\log(LR) \rightarrow \chi^2 \tag{3.4}$$

(Rost, 2004, S. 332)

In dem Programm ConQuest wird der doppelte negative Logarithmus der Likelihood unter „Final Deviance“ ausgegeben. Je kleiner der Deviance-Wert ausfällt, umso besser ist der *Fit* (die Passung) des angenommenen Modells auf die beobachteten Daten (Wu & Adams, 2006, S. 104). Damit ergibt sich wegen

$$\log(LR) = \log(L_0) - \log(L_1) \tag{3.5}$$

(Rost, 2004, S. 333)

[18] In diesem Zusammenhang wird auch von genesteten Modellen gesprochen (Osteen, 2010, S. 68).

dass die Differenz zweier von ConQuest ausgegebener Deviance-Werte direkt für den $\chi^2$-Test verwendet werden kann:

$$-2\log(LR) = -2\log(L_0) - (-2\log(L_1)) \tag{3.6}$$

Neben der Differenz der Deviance-Werte ist für den $\chi^2$-Test auch die Differenz der Parameteranzahl $t$ der beiden alternativen Modelle zu bestimmen. Aus der Differenz der Anzahl der Parameter ergibt sich die für die Testung benötige Anzahl an *Freiheitsgraden* (degrees of freedom, df) (Fischer, 1974, S. 360):

$$df = t(L_1) - t(L_0) \tag{3.7}$$

(in Anlehnung an Rost, 2004, S. 332)

Die Freiheitsgrade werden in ConQuest über „total number of estimated parameters“ ausgegeben (Wu et al., 2007, S. 125). Die Anzahl wird benötigt, um den kritischen Wert zu bestimmen, anhand dessen die Prüfgröße (die sich aus Gleichung 3.6 ergibt) gegen die $\chi^2$-Verteilung getestet wird. Der kritische Wert für die Prüfgröße ist der Wert, der nur mit einer Wahrscheinlichkeit von $p < 0,05$ überschritten wird (bei einem üblichen Signifikanzniveau von $\alpha = 0,05$) (Rost, 2004, S. 333). Dieser Wert ist aus einer $\chi^2$-Tabelle zu entnehmen. In Anlehnung an die Nullhypothese wird die identische Datenanpassung zwischen den beiden zu vergleichenden Modellen angenommen (Osteen, 2010, S. 68; de Gruijter & van der Kamp, 2008, S. 185; Borg & Staufenbiel, 2007, S. 364). Wenn die Prüfgröße den kritischen Wert übersteigt, also $> 0,05$ ist, so wird die Verträglichkeit des restriktiveren Modells (mit der Likelihood $L_0$) verworfen, da die Geltung unwahrscheinlicher als 5 Prozent ist (signifikantes Testergebnis). Gleichzeitig kann eine bessere Modellanpassung des allgemeineren Modells (mit der Likelihood $L_1$) angenommen werden. Liegt die Prüfgröße unterhalb der kritischen Grenze, so kann die Annahme beibehalten werden, dass das restriktivere Modell die beobachteten Daten erklärt (nicht signifikantes Ergebnis) (Carstensen, 2000, S. 160).[19]

### Informationstheoretische Maße

Anders als beim LR-Test können mit informationstheoretischen Maßen Modelle verglichen werden, bei denen ein Modell nicht das Obermodell des anderen sein muss (d. h., die Modelle müssen nicht genestet sein bzw. durch Restriktionen auseinander hervorgehen)

[19] Ein Spezialfall des Likelihoodquotiententests ist die Testung gegen das *saturierte* Modell. Ein saturiertes Modell erklärt die Daten perfekt, da es so viele Parameter spezifiziert wie es Antwortmuster gibt (Borg & Staufenbiel, 2007, S. 363). Es beschreibt die maximal mögliche Likelihood. Da es aber keine Annahmen über Zusammenhänge formuliert, ist seine Erklärungskraft nichtig. Die Freiheitsgrade werden nach $df = m^k - 1$ ermittelt, wobei $m$ die Anzahl der Antwortkategorien und $k$ die Anzahl der Items darstellt. Bei einem dichotomen Modell mit beispielsweise 15 Items ergeben sich bereits 32.767 Freiheitsgrade. Aufgrund der hohen Anzahl von Freiheitsgraden wird eine Anwendung eines Modelltests gegen das saturierte Modell nicht empfohlen (Bühner, 2011, S. 534).

| Informationsindex | Abkürzung | Berechnung |
|---|---|---|
| Akaike information criterion | AIC | -2 log L + 2 t |
| Bayes information criterion | BIC | -2 log L + (log n) t |
| Consistent AIC | CAIC | -2 log L + (log n) t + t |

Tabelle 3.2: Informationstheoretische Maße (in Anlehnung an Borg & Staufenbiel, 2007, S. 364)

(Bühner, 2011, S. 542; Borg & Staufenbiel, 2007, S. 364). Die Anforderung, dass sich die Modelle auf den gleichen Datensatz bzw. die gleiche Datenmatrix beziehen müssen, bleibt aber erhalten (s. o.). Ebenso werden ähnliche Informationen für den Vergleich herangezogen, die auch im LR-Test verwendet werden: die maximierte Modell-Likelihood und die Zahl der geschätzten Modellparameter. Zudem fließt der Stichprobenumfang teilweise mit in die Berechnung ein (Gollwitzer, 2008, S. 293; Rost, 2004, S. 339).

Gegenstand der informationstheoretischen Maße ist das Abwägen des log-Likelihoodwertes als Fitkriterium und der Parameterwerte als Einfachheitskriterium (Borg & Staufenbiel, 2007, S. 364). Das heißt, neben der Frage, welches Modell weniger von den Daten abweicht, wird auch geprüft, wie komplex ein Modell ist (Bühner, 2011, S. 541). Die Gewichtung von Likelihood und Parameterwerten unterscheidet die verschiedenen Informationskriterien. Allen gemein ist die Bevorzugung von Modellen mit hohem Modellfit, die zugleich einfach sind (Borg & Staufenbiel, 2007, S. 364).

Bei dem Vergleich von Informationsindizes zweier konkurrierender Modelle ist die Höhe der Abweichung, im Gegensatz zum LR-Test, kein Indikator dafür, ob ein Modell verworfen oder favorisiert werden sollte. Es kann unter mehreren konkurrierenden Modellen nur das relativ beste ausgewählt werden (Rost, 2004, S. 339). Bei der Wahl sind (neben dem Geltungs- und Einfachheitskriterium) ebenfalls inhaltlich-theoretische Überlegungen über die Angemessenheit und Brauchbarkeit zu berücksichtigen (Rost, 2004, S. 342). Insgesamt gilt: je kleiner der Wert der Informationsindizes, desto besser (Burnham & Anderson, 2004, S. 275). Die Komplexität eines Modells wird bei der Verrechnung von Likelihood $L$ (bzw. $-2 \log L$, der Deviance) und Parameterzahl t unterschiedlich „bestraft“, wie anhand von Tabelle 3.2 zu erkennen ist. Beim *Akaike information criterion* (AIC) hängt diese allein von der Anzahl $t$ zu schätzender Parameter des Modells ab. Bei der Berechnung des *Bayes information criterion* (BIC) und des *Consistent AIC* (CAIC) wird zusätzlich die Stichprobengröße $n$ berücksichtigt.

Die Berücksichtigung der Anzahl der Modellparameter, um Modelle mit vielen unnötigen Parametern zu bestrafen, ist ein wesentlicher Vorteil der Informationskriterien. Beim AIC erfolgt eine 1:1-Gewichtung von Parameteranzahl und log-Likelihood (da sie jeweils mit 2 gewichtet werden), während beim BIC und CAIC eine Gewichtung mit dem Logarithmus der Stichprobengröße stattfindet (vgl. Tabelle 3.2). Die Auswirkung der Parameteranzahl ist demnach beim BIC und CAIC (für $n \geq 8$) stärker als beim AIC. Bei einer Stichprobengröße

von beispielsweise 566 (wie beim gutschrift-Test) nimmt $\log n$ einen Wert von 6,34 an. Da ein komplexes Testmodell eine höhere Wahrscheinlichkeit hat, die empirischen Daten gut zu beschreiben, als ein einfaches, ist eine Gewichtung mit einer Funktion von $n$ bei großen Datensätzen laut Rost sinnvoll. Denn große Stichprobenumfänge erzeugen viele Antwortmuster, die zu modellieren sind, sowie weitere Modellparameter (Rost, 2004, S. 343). Komplexere Modelle werden mit dem BIC und CAIC stärker bestraft (Gollwitzer, 2008, S. 292 f.).

#### 3.3.2.2 Itemqualität

Neben der Bewertung der Güte des gesamten Modells mittels informationstheoretischer Maße, Likelihoodquotiententest und $\chi^2$-verteilter Prüfstatistik kann die Verträglichkeit der einzelnen Items eines Modells überprüft werden. Die Güte eines Items bzw. der *Itemfit* kann dabei anhand unterschiedlicher Charakteristiken bzw. Kennwerte einer Aufgabe eingeschätzt werden. Items, die keine gute Übereinstimmung mit dem Modell aufweisen, können von einem Test im Rahmen eines psychometrischen Analyseprozesses ausgeschlossen werden, um die Modellanpassung zu erhöhen. Die Anzahl eliminierter Items kann dabei durchaus hoch sein, weshalb empfohlen wird, zunächst mehr Items als angestrebt in einen Test aufzunehmen (Bortz & Döring, 2006, S. 227). Beim Itemreview handelt es sich um einen iterativen Prozess: Nachdem Items ausgeschlossen wurden, muss das Modell neu berechnet und im Anschluss die Charakteristiken aller im Modell verbliebenden Items erneut geprüft werden, da die Selektion Auswirkungen auf das ganze Modell hat (Wu & Adams, 2007, S. 69).

Neben der Aussortierung abweichender Items ist auch die abweichender Personen möglich. Während aber die Selektion von Aufgaben legitim ist, gilt diese von Personen als illegitim (Rost, 2004, S. 365). Items können mit Fehlern behaftet sein, sodass die Auswahl ggf. noch einmal überdacht und revidiert werden muss. Bei einer Veränderung der Stichprobe spricht Rost (2004, S. 365) von Annahmen der Datenmanipulation und Schönung von Ergebnissen. Unter speziellen Bedingungen gibt es jedoch Gründe, abweichende Personen zu identifizieren, die für die Analysen der vorliegenden Arbeit allerdings irrelevant sind und daher nicht dargestellt werden. Die Qualität eines Tests ist also abhängig von den Items, aus denen er besteht, und kann darüber verbessert werden. Im Gegensatz zur Veränderung einer Stichprobe haben sie zudem einen direkten Einfluss auf den Test. Die Itemanalyse, oder das Itemreview, ist daher zentrales Mittel der Testkonstruktion und -auswertung, bei dem problematische Items identifiziert und eliminiert werden (Bortz & Döring, 2006, S. 217). Für Itemcharakteristikanalysen sollte die Stichprobe ausreichend groß sein. Hierzu existieren keine Standards. Embretson und Reise (zitiert n. Osteen, 2010, S. 68) halten 500 Personen für empfehlenswert, während sie Parameterschätzungen bei weniger als 350 Personen als unstabil und vorsichtig zu interpretieren benennen. Bei komplexer werdenden, mehrparametrigen Modellen sollten auch die Personenstichprobengrößen ansteigen (Osteen, 2010, S. 68).

In Anlehnung an PIRLS und PISA werden die Itemschwierigkeitsindizes, die Diskriminierungsstatistiken und der *Infit mean square* in Rahmen des Itemreviewprozesses betrachtet.[20] Als *dodgy items* werden die modellinkonformen Aufgaben bezeichnet, deren Schwierigkeitsgrad unpassend für die Stichprobe ist, deren Trennschärfe klein ausfällt oder deren Itemfit außerhalb definierter Grenzen liegt (Martin, Kennedy & Trong, 2007, S. 136; OECD, 2005, S. 127; Adams, 2002, S. 105). Die psychometrischen Kriterien für die Qualitätsbeurteilung und Überprüfung der Items werden nachfolgend ausgeführt.

### Itemschwierigkeit

Eine möglichst breite Schwierigkeitsstreuung ist bei einem Test wichtig, um damit die Fähigkeit aller Personen einer Stichprobe gut beschreiben zu können. Das heißt, ein Test sollte leichte und schwierige Items enthalten (Bortz & Döring, 2006, S. 218). Leichte Items lösen fast alle Personen und die Lösungswahrscheinlichkeit fällt hoch aus, während schwierige Items nur von einer kleinen Personengruppe richtig beantwortet werden. Enthält ein Test ausschließlich sehr leichte oder sehr schwierige Items, ist dieser wenig informativ, da er zwischen leistungsstarken und -schwachen Personen unzureichend differenziert (Wu & Adams, 2007, S. 68 f.). Um Personenunterschiede deutlich zu machen, sollte ein möglichst breites Fähigkeitsspektrum mit einem Test erfasst werden (Wright & Stone, 1979, S. 4 ff.). Die Analyse der Aufgabenschwierigkeiten ist daher ein wichtiges Mittel für einen Abgleich zwischen erwarteter und beobachteter Schwierigkeit.

Die Schwierigkeit jedes Items ist in ConQuest in der Spalte „Estimate“ aufgeführt. Zudem vermittelt die *Wright Map*[21] einen guten visuellen Eindruck über die Verteilung und Passung der Personenfähigkeiten und Itemschwierigkeiten (Bond & Fox, 2007, S. 54 ff.). Wenn die Verteilung der Personenfähigkeiten mit der Lage der Itemschwierigkeiten entlang der Logitskala übereinstimmt, dann passt der Test gut auf die Personengruppe. Ist die Verteilung linksschief (bzw. rechtssteil), zeigt dies, dass der Test eher zu leicht konstruiert worden ist. Der Schwierigkeitsgrad ist klein, da die mittlere Lösungswahrscheinlichkeit über 50 Prozent liegt, also ein großer Teil der Stichprobe über die Hälfte der Aufgaben richtig löst (Kelava & Moosbrugger, 2008, S. 92). Eine rechtsschiefe Verteilung (linkssteil) zeigt hingegen einen zu schweren Test an, da ein großer Teil der Stichprobe weniger als die Hälfte der Aufgaben richtig löst.

[20] Die *T-Statistik* (nähere Informationen dazu finden sich z. B. in Wright und Masters (1982, S. 101 ff.)) wird beim Itemreviewprozesses nicht als striktes Selektionskriterium für Items verwendet. Bei der Itemkontrolle wurde der Wert zwar stets berücksichtigt, nahm aber ausschließlich dann abweichende Werte an, wenn sich auch der Infit mean square außerhalb der definierten Grenzen befand. Diese Beobachtung passt zu den Empfehlungen von Bond und Fox (2007, S. 241 f.) sowie Wilson (2005, S. 129), Aufgaben mit problematischem T-Wert ausschließlich dann zu eliminieren, wenn auch der Infit schlecht ausfällt.

[21] In Abschnitt 4.2 wird die Wright Map für beide Rechtschreibtests visualisiert.

### Klassische Itemdiskrimination

Die klassische[22] Trennschärfe oder klassische Diskrimination eines Items ergibt sich aus der Korrelation zwischen den Beantwortungen des Items und den aus allen restlichen Aufgabenwerten gebildeten Gesamtscores. Die Trennschärfe drückt aus, wie gut eine Aufgabe zwischen Personen mit verschiedenen (z. B. hohen und niedrigen) Fähigkeitsausprägungen zu trennen vermag (Wu & Adams, 2007, S. 64). Sie wird in ConQuest als Produkt-Moment-Korrelation angegeben (Wu et al., 2007, S. 149 f.). Die Trennschärfe ist bei dem dichotomen Raschmodell mit der punktbiseralen Korrelation[23] (pointbiserial index of discrimination, Pt Bis) identisch (Wu et al., 2007, S. 150) und umfasst einen korrelationstypischen Wertebereich von $-1$ bis $+1$ (Bortz & Döring, 2006, S. 220). Items mit negativer Trennschärfe und nahe dem Wert 0 sollten aus einem Test ausgeschlossen werden (Kelava & Moosbrugger, 2008, S. 85). Je höher der Diskriminationskoeffizient ausfällt, desto stärker ist der Zusammenhang zwischen den Aufgabenscores von Personen und ihren Gesamttestwerten. Zu beachten gilt dabei allerdings, dass die Trennschärfe von der Schwierigkeit eines Items abhängt (Bortz & Döring, 2006, S. 220). Ist ein Item besonders leicht oder schwer, so sinkt sie. Wenn ein Test aber sowohl die Fähigkeiten besonders leistungsstarker als auch -schwacher Schülerinnen und Schüler berücksichtigen möchte, um auch die Randbereiche eines Konstrukts differenziert erfassen zu können, so sind Trennschärfeeinbußen in Kauf zu nehmen (Bortz & Döring, 2006, S. 220; Rost, 2004, S. 99).

Bei der Itemanalyse wurde in der hier vorliegenden Arbeit daher darauf geachtet (analog z. B. zu dem Vorgehen bei der Pilotierung der Bildungsstandards (Winkelmann & Böhme, 2009, S. 37)), Aufgaben mit extremen Schwierigkeitsgraden nicht überproportional aus dem Test zu entfernen. In der Literatur finden sich unterschiedliche Vorgaben für Diskriminationsgrenzen, die anzustreben sind. Bortz und Döring (2006, S. 220) beschreiben Itemtrennschärfen zwischen $0,3$ und $0,5$ als mittelmäßig und größer $0,5$ als hoch. In PISA wird ein Index bei der Itemselektion von $\geq 0,25$ angegeben (OECD, 2005, S. 123; Adams, 2002, S. 102). Da aber auch besonders einfache und schwierige Aufgaben bei den Analysen in Kapitel 4 berücksichtigt werden sollen, wurde ein leicht niedrigerer Grenzwert von $0,2$ festgelegt, der auch bei PIRLS (Martin et al., 2007, S. 136) sowie Wu und Adams (2007, S. 64) angegeben wird.

### Item response model fit

Der Modellfit eines Items ist ein Indikator für die Kompatibilität der Daten mit dem Raschmodell bzw. ein Indikator für die Dimensionalität der Testaufgaben (Linacre, 2002, S. 878). Es wird geprüft, ob ein Item zu den mit dem Modell verbundenen Annahmen passt,

[22] Der Begriff *klassische* Trennschärfe wird hier verwendet, da sie hier analog zur Berechnung im Rahmen der KTT genutzt wird (Rost, 2004, S. 369).

[23] Mit der punktbiserialen Korrelation kann der Zusammenhang einer intervallskalierten und einer dichotomen Variable errechnet werden (B. Rasch, Friese, Hofmann & Naumann, 2004, S. 124 ff.).

also raschskalierbar ist. Die Modellverträglichkeit wird über Residuen-basierte Fitmaße ermittelt. Als *Residuum* $Y_{in}$ wird die Abweichung zwischen beobachteter Reaktion $X$ und erwarteter Reaktion $E$ einer Person $n$ auf ein Item $i$ bezeichnet (Bond & Fox, 2007, S. 237 f.; Wilson, 2005, S. 127; Embretson & Reise, 2000, S. 235):

$$Y_{in} = X_{in} - E_{in} \tag{3.8}$$

(Wilson, 2005, S. 127)

Der Fit $u_i$ eines Items $i$ setzt sich aus der Summe aller quadrierten standardisierten[24] Residualgrößen $Z_{in}^2$ über alle Personen in Bezug auf ein bestimmtes Item zusammen, die durch die Personenanzahl $N$ dividiert wird, woraus sich der *unweighted meansquare* (MNSQ, Abweichungsquadrate) ergibt:

$$u_i = \frac{\sum_{n=1}^{N} Z_{in}^2}{N} \tag{3.9}$$

(in Anlehnung an Wright & Masters, 1982, S. 99)

Diese Itemfitstatistik reagiert sehr sensibel auf Ausreißer, also auf vereinzelte und extreme Antworten von Personen, für die das Item unerwartet zu leicht oder zu schwer ist, sodass das Item als nicht modellkonform angenommen wird (Linacre, 2002, S. 878; Carstensen, 2000, S. 177; Wright & Masters, 1982, S. 99). Daher wird die in Gleichung 3.9 dargestellte Statistik auch *Outfit*-Statistik genannt. Um dieses Problem zu beheben, haben Wright und Masters die Residuen gewichtet:

$$v_i = \frac{\sum_{n=1}^{N} Z_{in}^2 W_{in}}{\sum_{n=1}^{N} W_{in}} = \frac{\sum_{n=1}^{N} Y_{in}^2}{\sum_{n=1}^{N} W_{in}} \tag{3.10}$$

(in Anlehnung an Wright & Masters, 1982, S. 99)

Beim *weighted MNSQ* (also WMNSQ oder auch *Infit*) wird jedes quadrierte standardisierte Residuum mit seiner jeweiligen Varianz $W_{in}$ gewichtet. Die Varianz fällt kleiner für Personen mit unerwarteten Reaktionen auf Item $i$ aus, womit sich der Einfluss ihrer Antworten auf $v_i$ reduziert. Im umgekehrten Fall fällt die Varianz bei hoher Übereinstimmung zwischen den Personen- und Itemparametern groß aus. Dadurch werden extreme Abweichungen zwischen den erwarteten und beobachteten Reaktionen von Personen auf Items abgeschwächt. Das heißt, die erwarteten Antworten von Personen wirken sich stärker auf das Ergebnis aus als die unerwarteten (Wright & Masters, 1982, S. 101). Der Infit sollte daher anstatt des Outfits für die Itemanalyse verwendet werden (Bond & Fox, 2007, S. 286).

Der (W)MNSQ hat einen Erwartungswert von 1 (Bond & Fox, 2007, S. 289). Werte von 1 weisen demnach einen perfekten Fit auf, d. h., die Aufgabe passt perfekt zu dem Modell. Aufgaben mit einem Infit nahe 1 gelten als gut mit dem Modell verträglich. Sie

[24] Für die Standardisierung fließt die Varianz $W_{in}$ ein: $Z_{in} = \frac{Y_{in}}{W_{in}^{1/2}}$ (Wright & Masters, 1982, S. 98 f.).

werden angestrebt und deuten auf eine angemessene Passung zwischen prognostizierten und beobachteten Reaktionen hin. Ein Wert von beispielsweise 1,15 (1,00 + 0,15) bedeutet, dass 15 Prozent mehr beobachtete Variation in den Antwortreaktionen des Items vorliegt, als von dem Modell prognostiziert (Osteen, 2010, S. 68; Bond & Fox, 2007, S. 239 ff.). Ein Datenbeispiel für die graphische Veranschaulichung des Itemfits mittels ICC erfolgt in Abschnitt 4.2.2.

Es gibt unterschiedliche Empfehlungen über die angestrebte Ober- und Untergrenze des WMNSQ (Bond & Fox, 2007, S. 286). Adams und Khoo (1996, zitiert n. Wilson, 2005, S. 129) geben beispielsweise eine Spanne von 0,75 bis 1,33 sowie Linacre (2002, S. 878) ein Intervall zwischen 0,5 und 1,5 an. Bei den in Kapitel 4 dargestellten Analysen wurde die Abweichung vom Erwartungswert des WMNSQ in Orientierung an die Large-Scale-Assessments, wie z. B. PISA (OECD, 2012, S. 138), getroffen, bei denen der Wert nicht niedriger als 0,8 und nicht höher als 1,2 ausfallen sollte.

### 3.3.3 Latente Profil- und Klassenanalyse

Unter dem Begriff der probabilistischen Clusterverfahren wird die Analyse latenter Klassen (*latent class analysis*, LCA) und latenter Profile (*latent profile analysis*, LPA) subsumiert. Es handelt sich dabei um modellbasierte Verfahren zur empirischen Klassifikation (McCutcheon, 1987, S. 7). Grundlegender Gedanke dieser Verfahren ist es, Unterschiede von Personen[25] im beobachteten Antwortmuster oder -verhalten unter der Annahme von latenten Klassen zu identifizieren (J. Wang & X. Wang, 2012, S. 290; Magidson & Vermunt, 2004, S. 178). Dabei wird von heterogenen Stichproben ausgegangen, die sich aufgrund latenter kategorialer Variablen unterscheiden. In den latenten Klassen haben die latenten Variablen dann gleiche Ausprägungen und die Personen befinden sich in homogenen Subgruppen (Krauth, 1983, S. 351). Ziel probabilistischer Clusteranalysen ist die Maximierung der Einteilung von Personen anhand ihrer Itemreaktionen in homogene latente Klassen bei möglichst hoher Heterogenität zwischen den voneinander abgegrenzten Klassen (Bacher et al., 2010, S. 16 ff.; Bortz & Schuster, 2010, S. 453). Anders als bei der Faktorenanalyse handelt es sich u. a. also um Personen-zentrierte und nicht um Variablen-zentrierte Ansätze (J. Wang & X. Wang, 2012, S. 291; Collins & Lanza, 2010, S. 8).

Die LPA (Vermunt & Magidson, 2002, S. 90 ff.; Lazarsfeld & Henry, 1968, S. 228 ff.) basiert auf der LCA, welche bereits 1950 erstmalig von Lazarsfeld vorgestellt und 1968 von Lazarsfeld und Henry sowie 1974 von Goodman weiterentwickelt wurde. Während die LCA allerdings mit dichotomen und polytomen manifesten Variablen arbeitet, nutzt die LPA kontinuierliche Indikatorvariablen (L. K. Muthén & B. O. Muthén, 2012, S. 170; Lubke & B. Muthén, 2005, S. 23). Das Verfahren der LPA wird in Abschnitt 4.4.2 genutzt, um latente Klassen über die individuellen Leistungswerte der Schülerinnen und Schüler in den beiden Rechtschreibtests zu bilden.

[25] Es kann sich – neben Personen – auch um Aggregate (Organisationen, Länder etc.) oder andere konkrete Entitäten handeln (Bacher, Pöge & Wenzig, 2010, S. 15).

### Modellgleichung

Im Folgenden sollen die zentralen Grundannahmen des allgemeinen kategorialen Testmodells in Orientierung an Rost (2004) dargestellt werden. Die Basis bildet dabei ein dichotomes LCA-Modell, da es die zentralen Ideen umfasst und Modelle mit anders skalierten Daten aus diesem abgeleitet werden. Einige der folgenden Annahmen sind bereits aus dem Raschmodell bekannt, sollen aber dennoch in diesem Rahmen ausgeführt werden, da sie hier speziell auf die Verfahren latenter Klassenanalysen angewendet werden und in diesen spezifischen Gesamtzusammenhang eingebettet sind.

I. Eine erste Annahme drückt aus, dass jede Person einer Klasse eine spezifische und *konstante* Wahrscheinlichkeit für das Antwortverhalten auf ein Item (z. B. für das Lösen oder nicht-Lösen eines Items) besitzt (Rost & Langeheine, 1997, S. 28):

$$p(X_{vi} = x \mid \theta_v = g) = \pi_{ig}^{x}(1 - \pi_{ig})^{1-x} \tag{3.11}$$

(Rost, 2004, S. 158)

Bei $X_{vi}$ handelt es sich um die Antwort einer Person $v$ auf ein Item $i$. Die Wahrscheinlichkeit $p$ der Itemantwort $x \in \{0, 1\}$ ist abhängig von der klassifizierten Fähigkeit $\theta_v$ der Person. Bei $\pi$ handelt es sich um einen Wahrscheinlichkeitsparameter (Rost, 2004, S. 158). Dieser kann nur Werte zwischen 0 und 1 annehmen. $\pi_{ig}$ steht für die (konstante) Lösungswahrscheinlichkeit von Item $i$ in Klasse $g$. Anhand von Gleichung 3.11 lässt sich damit erkennen, dass die Lösungswahrscheinlichkeit[26] eines Items einer Person ausschließlich von deren Klassenzugehörigkeit abhängt (Gollwitzer, 2008, S. 284).

II. Die zweite Annahme besagt, dass die Klassen disjunkt und exhaustiv sind. Das heißt, jede Person wird einer Klasse, aber genau auch nur einer Klasse zugeordnet (Collins & Lanza, 2010, S. 39 ff.). Dies impliziert, dass sich die relativen Klassengrößen zu 1 aufsummieren (Gollwitzer, 2008, S. 286; McCutcheon, 1987, S. 19):

$$\sum_{g=1}^{G} \pi_g = 1 \tag{3.12}$$

(Rost, 2004, S. 158)

Die relative Klassengröße $\pi_g$, also die Wahrscheinlichkeit, dass eine Person einer Klasse $g$ angehört, ist dabei unbekannt. Die Anzahl der Klassen $G$ ist zwar ebenfalls unbekannt, aber kein zu schätzender Modellparameter. Die Klassenanzahl muss daher über Modellvergleiche und mit Hilfe von Gütekriterien (s. u.) ermittelt werden (Geiser, 2011, S. 237).

Mit den beiden zuvor vorgestellten Parametern $\pi_{ig}$ und $\pi_g$ lässt sich die unbedingte (von der jeweiligen Klasse unabhängige) Lösungswahrscheinlichkeit einer Person $v$

[26] Lösungswahrscheinlichkeit bezeichnet den Spezialfall der Antwortwahrscheinlichkeit bei dichotomen Items.

für eine Aufgabe $i$ darstellen. Dafür wird über alle Klassen die Summe der Produkte der Klassengrößenwahrscheinlichkeiten und der bedingten (also von der jeweiligen Klasse abhängigen) Lösungswahrscheinlichkeiten eines Items berechnet:

$$p(X_{vi} = 1) = \sum_{g=1}^{G} \pi_g \pi_{ig} \tag{3.13}$$

(Rost, 2004, S. 159)

III. Eine dritte Annahme geht von Itemhomogenität aus. Das heißt, alle Items messen dieselbe Personenvariable (Rost, 2004, S. 158 f.). Die folgende Gleichung stellt diesen Sachverhalt dar:

$$p(\vec{x}) = \sum_{g=1}^{G} \pi_g \, p(\vec{x} \mid g) \tag{3.14}$$

(in Anlehnung an Rost, 2004, S. 159)

Die Antwortmuster oder Antwortpattern sind mit $\vec{x}$ dargestellt. Aus der Annahme der Itemhomogenität wird gefolgert, dass sich die unbedingten Patternwahrscheinlichkeiten $p(\vec{x})$ auf die gleiche Weise berechnen lassen wie zuvor die unbedingten Lösungswahrscheinlichkeiten (Rost, 2004, S. 159).

IV. Die vierte Annahme umfasst das Prinzip der lokalen stochastischen Unabhängigkeit und ist zentral für das allgemeine kategoriale Testmodell (Rost, 2006, S. 280; Lazarsfeld & Henry, 1968, S. 22). Die Reaktion einer Person auf ein Item ist damit ausschließlich durch ihre Klassenzugehörigkeit bestimmt (Collins & Lanza, 2010, S. 46). Das bedeutet, dass die beobachteten Zusammenhänge zwischen den manifesten Variablen ausschließlich auf die latente Variable zurückzuführen sind. Wenn also die latente Variable konstant gehalten wird und die manifesten Variablen in den Klassen getrennt untersucht werden, verschwinden die Zusammenhänge: „The criterion of local independence, then, provides a method for determining whether relationships among a set of observed measures are due to some unmeasured explanatory variable. When a set of interrelated variables are found to be locally independent within categories of some additional variable, we say that the additional variable ‚explains' the observed relationships – that the additional variable represents the true variable interest, and that once it is considered all of the other measures are unrelated." (McCutcheon, 1987, S. 16)

Die bedingten Patternwahrscheinlichkeiten in einer Klasse $p(\vec{x} \mid g)$ (also für ein Antwortmuster $\vec{x}$ unter der Bedingung der Klassenzugehörigkeit zu $g$) können durch die lokale stochastische Unabhängigkeit als Produkt der einzelnen Lösungswahrscheinlichkeiten der Items in einem Test mit $k$ Items ausgedrückt werden:

$$p(\vec{x} \mid g) = \prod_{i=1}^{k} \pi_{ig}^{x_i} (1 - \pi_{ig})^{1-x_i} \tag{3.15}$$

(in Anlehnung an Rost, 2004, S. 159)

Aus den Gleichungen 3.14 und 3.15 ergibt sich nach Lazarsfeld (1950, zitiert n. Rost, 2004, S. 160) folgendes *Grundmodell* für alle Testmodelle mit kategorialer Personenvariable – das Latent-Class-Modell:

$$p(\vec{x}) = \sum_{g=1}^{G} \pi_g \prod_{i=1}^{k} \pi_{ig}^{x_i} (1 - \pi_{ig})^{1-x_i} \tag{3.16}$$

(in Anlehnung an Rost, 2004, S. 159)

Aus den Modellannahmen ist ersichtlich, dass die Zuordnung von Personen zu Klassen über Wahrscheinlichkeitsaussagen erfolgt. Wie groß die Wahrscheinlichkeit ist, dass eine Person mit einem bestimmten Antwortmuster $\vec{x}$ einer Klasse $g$ zugeordnet ist, wird über die bedingte Klassenwahrscheinlichkeit ausgesagt. Sie lässt sich mit Hilfe des Bayes-Theorems wie folgt bestimmen (Gollwitzer, 2008, S. 286):

$$p(g \mid \vec{x}) = \frac{\pi_g \, p(\vec{x} \mid g)}{\sum_{h=1}^{G} \pi_h \, p(\vec{x} \mid h)} \tag{3.17}$$

(in Anlehnung an Rost, 2004, S. 160)

Die bedingte Klassenwahrscheinlichkeit ergibt sich aus den Klassenzuordnungswahrscheinlichkeiten $\pi_g$ multipliziert mit der bedingten Patternwahrscheinlichkeit $p(\vec{x} \mid g)$ (Antwortmuster einer Person aus einer Klasse) dividiert durch die Summe dieser Produkte über alle Klassen. Auf diese Weise kann für jede Person mit einem spezifischen Antwortmuster eine Wahrscheinlichkeit für die Klassenzugehörigkeit berechnet werden. Letztendlich wird eine Person der Klasse zugeordnet, für die sie die höchste Zuordnungswahrscheinlichkeit besitzt (Rost, 2004, S. 160). Durch diese nicht manifeste Zuordnung sind bei den probabilistischen Clusterverfahren Messfehler berücksichtigt, da sie vom Modell durch die Wahrscheinlichkeitsparameter impliziert werden (Rost, 2004, S. 156).

### Verhältnis zum Raschmodell

Wie unter den grundlegenden Modellannahmen erläutert, wird die Wahrscheinlichkeit einer Klassenzugehörigkeit über das Antwortverhalten bzw. -muster berechnet. Jede Person hat für jede Klasse eine spezifische Wahrscheinlichkeit, d. h., sie kann einer Klasse mehr oder weniger stark angehören (Rost, 2006, S. 278). Die Zuordnung erfolgt also nicht deterministisch, sondern probabilistisch. Die LCA/LPA ist also ebenso wie das Raschmodell ein Modell der Item-Response-Theorie (Gollwitzer, 2008, S. 281).

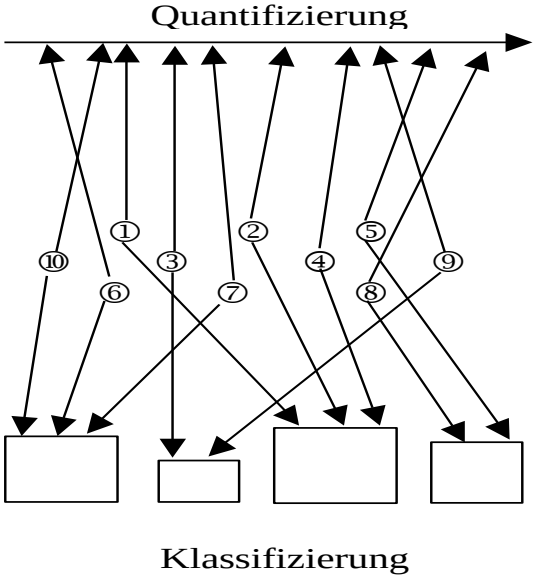



Abbildung 3.11: Anordnung von Personen auf einem Kontinuum oder Zuordnung zu Klassen (Rost, 2004, S. 148)

Abgrenzend zum Raschmodell ergibt sich die Klassenzugehörigkeit bei den probabilistischen Clusteranalysen nicht aus der Anzahl der richtigen Antworten oder gelösten Items, sondern aus dem Muster. Es wird nicht danach gefragt „Wie viele Items wurden richtig beantwortet?“, sondern „Welche Items wurden richtig beantwortet?“. Beim Raschmodell ist die latente Variable quantitativ und die Personen werden entlang eines kontinuierlichen latenten Kontinuums abgebildet. Personen mit unterschiedlichen Eigenschaftsausprägungen können in eine Rangreihenfolge gebracht werden. Bei latenten Klassenmodellen werden die Personen hingegen nicht quantifiziert, sondern klassifiziert. Die latente Variable ist qualitativ und Personen werden kategorial in verschiedene Klassen eingeteilt (Rost, 2004, S. 148). Diese unterschiedlichen Ziele der quantitativen und qualitativen Testanalyse sind in Abbildung 3.11 visualisiert. Bei der LCA/LPA werden dementsprechend die Antwortmuster auch nicht auf einer ICC abgetragen, sondern in Form von Itemprofilen (gegenüber Itemfunktionen). Auf der Abszisse sind dann die qualitativen Indikatoren, auf der Ordinate bei der LCA die Lösungswahrscheinlichkeiten und bei der LPA die klassenspezifischen Itemmittelwerte abzulesen. Itemprofile geben damit Auskunft über Lösungswahrscheinlichkeit der Items bzw. die Mittelwerte in den Klassen.

Verglichen mit dem Raschmodell sind die LC-Modelle weniger restiktiv bezüglich der Voraussetzung auf Personenhomogenität (Rost & Langeheine, 1997, S. 29). Mit Personenhomogenität ist gemeint, dass für alle Personen eine homogene Eigenschaft gemessen wird, da die Schätzungen der Itemparameter für jede beliebige (raschkonforme) Teilstichprobe gleich ausfallen (spezifische Objektivität) (Bühner, 2011, S. 539). Bei der LCA kann eine Klasse im Vergleich zu einer weiteren Klasse innerhalb des Modells sehr unterschiedliche Itemparameter besitzen. Auf der anderen Seite ist das LC-Modell – im Hinblick darauf, dass alle Unterschiede zwischen den Personen mit einer relativ kleinen Anzahl an Klassen

erklärt werden müssen – restriktiver als das Raschmodell. Alle Personen einer Klasse werden bezüglich ihres Antwortverhaltens identisch behandelt, während beim Raschmodell jede Person mit eigener Position auf dem latenten Kontinuum verortet ist (vgl. Abbildung 3.11) (Rost & Langeheine, 1997, S. 29).

Rost (2004, S. 162) spricht im Kontext der LCA von „full-information methods“, da bei den Klassenberechnungen alle in den Daten enthaltenen Informationen genutzt werden. Beim Raschmodell greift das Prinzip der suffizienten Statistiken: unterschiedliche Antwortmuster werden nicht voneinander unterschieden (vgl. Abschnitt 3.3.2). Bei der LCA erfolgt eine Unterscheidung; hat ein Test beispielsweise 5 dichotome Items, dann gäbe es $2^5 = 32$ theoretisch mögliche Pattern mit potenziell verschiedenen Häufigkeiten. Diese Häufigkeiten der Antwortmuster werden in einer Stichprobe ausgezählt und bilden die Datenbasis für die Parameterschätzungen.

### Parameterschätzung

Ähnlich wie bei der Parameterschätzung beim Raschmodell (vgl. Abschnitt 3.3.2) wird auch bei der Clusteranalyse die Likelihood mittels Maximum-Likelihood-Methode optimiert. Die zu schätzenden Modellparameter sind dabei der Klassengrößeparameter $\pi_g$ und die klassenspezifischen Lösungswahrscheinlichkeiten $\pi_{ig}$ (vgl. Gleichung 3.13) (Rost, 2004, S. 318 f.). Auch hier werden die unbekannten Modellparameter iterativ geschätzt (Gollwitzer, 2008, S. 289). Dafür wird der *Expectation-Maximization-Algorithmus*[27] (EM-Algorithmus, im Deutschen manchmal auch Erwartungswert-Maximierungs-Algorithmus genannt) verwendet, der auf Goodman (1974) zurückgeht (Bacher et al., 2010, S. 352; Formann, 2010, S. 557; McCutcheon, 1987, S. 21 ff.).[28] Dabei werden durch Maximierung der Likelihoodfunktion die Personen schrittweise derselben latenten Klasse zugeordnet, die dieselben Antwortwahrscheinlichkeiten bezüglich der Items haben (Rost, 2004, S. 156; B. Muthén, 2001a, S. 7).

Es kann beim Modellschätzprozess zu nicht unüblichen Problemen kommen, wenn ein lokales Maximum, das nicht gleichzeitig globales Maximum ist, auftritt. Das heißt, dass bei dem iterativen Berechnungsprozess keine optimale Lösung mit der größten Wahrscheinlichkeit für eine zuverlässige Datenreproduktion durch die Modellparameter gefunden wird (Collins & Lanza, 2010, S. 89 ff.). Daher bietet das in dieser Arbeit eingesetzte Programm „Mplus“ die Möglichkeit, multiple Startwerte festzulegen. So kann der Anwender die Anzahl von zufälligen Startwerten verändern und erhöhen, damit die optimale Lösung mit dem größten Likelihoodwert mit hoher Wahrscheinlichkeit gefunden wird (L. K. Muthén & B. O. Muthén, 2012, S. 465 ff.).

[27] Für ausführliche Informationen zur Maximum-Likelihood-Methode sowie zum EM-Algorithmus in LCA siehe beispielsweise McCutcheon (1987, S. 21 ff.) oder Goodman (1974).

[28] Im Gegensatz zu der klassischen Clusteranalyse müssen daher im Vornherein nicht das Ähnlichkeits- und Distanzmaß, der Algorithmus der Berechnung sowie die Distanz zwischen den Clustern bestimmt werden. Diese Festlegungen sind nach Rost (2004, S. 156) „eine relativ willkürliche Auswahl unter mehreren Alternativen“. Damit haben probabilistische Clusterverfahren einen Vorteil gegenüber deterministischen Verfahren.

### Modellgütebeurteilung

Die Anzahl der latenten Klassen kann nicht direkt berechnet werden[29] (Gollwitzer, 2008, S. 282; Rost, 2004, S. 158). In der Regel werden daher a priori mehrere Modelle mit unterschiedlichen Klassenanzahlen berechnet und im Anschluss gegenübergestellt (Formann, 2010, S. 557). Im Vergleich zur deterministischen Clusteranalyse können bei der probabilistischen Clusteranalyse formal bessere Statistiken zur Beurteilung der Modellgüte herangezogen werden (Bacher et al., 2010, S. 20, 353; Magidson & Vermunt, 2004, S. 176; Vermunt & Magidson, 2002, S. 90). Dabei wird der relative Modellfit betrachtet, d. h., es wird eine Auswahl zwischen zwei benachbarten und zugleich konkurrierenden Modellen getroffen. Die jeweils geschätzte $k$-Klassenlösung wird der Klassenlösung mit $k-1$ Klassen gegenübergestellt.

Für die Modellwahl stehen informationstheoretische Maße und Likelihood-basierte Tests zur Verfügung. Die informationstheoretischen Maße sind AIC, BIC und CAIC (vgl. Abschnitt 3.3.2.1) sowie der *sample-size-adjusted BIC* (ssa BIC). Beim ssa BIC wird die Stichprobengröße ersetzt durch:

$$n* = \frac{(n+2)}{24}$$

(in Anlehnung an Yang, 2006, S. 1093)

Durch diese Modifikation fällt die Bestrafung durch zusätzliche Parameter geringer aus als beim BIC (Yang, 2006, S. 1093).

Unter den Likelihood-basierten Tests finden der *Vuong-Lo-Mendell-Rubin Likelihood-Ratio-Test* (VLMRT), der *Lo-Mendell-Rubin Adjusted-Likelihood-Ratio-Test* (LMRAT) (Lo, Mendell & Rubin, 2001) sowie der *Bootstrap-Likelihood-Ratio-Differenzentest*[30] (BLRT) Anwendung (McLachlan & Peel, 2000, S. 192 ff.). Über diese sind inferenzstatistische Aussagen möglich, jedoch sind sie im Vergleich zu den informationstheoretischen Maßen weniger erprobt und die „Güte [...] noch nicht vollständig überprüft." (Nussbeck, Eid & Geiser, 2010, S. 567) Bei den Tests zeigt ein signifikanter Wert ($p \leq 0{,}05$) an, dass das $k-1$-Klassemodell abgelehnt und das $k$-Klassenmodell angenommen werden sollte (Geiser, 2011, S. 265 ff.): „The p-value obtained represents the probability that the data have been generated by the model with one less class. A low p-value indicates that the model with one less class is rejected in favor of the estimated model." (L. K. Muthén & B. O. Muthén, 2012, S. 738) Der Likelihood-Ratio-Chi-Quadrat-Test ist nicht geeignet, eine Entscheidung über $k$ und $k-1$ Klassen zu treffen, da die Differenzen der Likelihoodwerte üblicherweise nicht $\chi^2$-verteilt sind (Samuelson & Raczynski, 2013, S. 309; Nylund et al., 2007, S. 542 ff.; Magidson & Vermunt, 2004, S. 176). VLMRT, LMRAT

[29] Falls doch bereits (Vor-)Annahmen über die Klasseanzahl getroffen werden können, so wird auch von konfirmatorischen (im Gegensatz zu explorativen) Clusteranalysen gesprochen (Bacher et al., 2010, S. 22 f.; McCutcheon, 1987, S. 37 ff.).

[30] Für nähere Informationen zur Bootstrapping-Prozedur siehe Langeheine, Pannekoek und van de Pol (1996) sowie Nylund, Asparouhov und Muthén (2007, S. 543 f.).

und BLRT nutzen eine abgeleitete Verteilung und korrigieren damit die Verteilung der log-Likelihood-Differenzen (J. Wang & X. Wang, 2012, S. 293).

Von Bedeutung für die Modellgüte ist auch die Genauigkeit der Klassifikation. Der im Rahmen der LCA/LPA von Mplus berechnete *Entropy-Wert* gibt an, wie klar unterscheidbar die Klassen sind (Celeux & Soromenho, 1996, S. 200). Werte nahe 1 zeigen eine deutliche Abgrenzung voneinander an (Collins & Lanza, 2010, S. 74 f.; Dias & Vermunt, 2006). Als Maß für die Sicherheit der Klassifikation werden gegenüber der Entropy häufiger die mittleren Klassenzuordnungswahrscheinlichkeiten betrachtet (Geiser, 2011, S. 249). Sie geben im Vergleich keine globale Modellinformation, sondern differenzierte Hinweise auf die Stärke der richtigen Zuordnung der Personen zu der Klasse für jede einzelne Klasse. Nach Rost (2006, S. 278) sollten sie $> 0,8$ betragen. Auch hier beinhalten Werte nahe 1 eine hohe Sicherheit der Klassenzuordnung (Geiser, 2011, S. 250).

Neben den Ergebnissen der oben genannten Tests und Kriterien sind auch inhaltliche und theoriebasierte Überlegungen bei der Wahl des Klassenmodells und der Interpretation der Lösungen zu berücksichtigen (J. Wang & X. Wang, 2012, S. 295; Geiser, 2011, S. 271). Dies gilt insbesondere vor dem Hintergrund, dass nicht das eine eindeutige und entscheidende statistische Maß für eine gültige Klassenwahl existiert: „To date, there is not common acceptance of the best criteria for determining the number of classes in mixture modeling, despite various suggestions." (Nylund et al., 2007, S. 537). In einer Simulationsstudie von Nylund et al. (2007) wurde daher die Performance der Modellgütetests LRT, VLMRT, LMRAT und BLRT sowie der Informationsindizes AIC, CAIC, BIC und ssa BIC im Rahmen von LCAs miteinander verglichen. Als Ergebnis konnte festgehalten werden, dass bei den Likelihood-basierten-Tests der BLRT und bei den informationstheoretischen Maßen der BIC als gute Indikatoren genutzt werden können, um zuverlässige Klassenentscheidungen zu treffen. Dies wird bei den Analysen in Abschnitt 4.4.2 entsprechend berücksichtigt.

### Latente Profilanalysen in Large-Scale-Assessments

LPA werden auch in den großen Schulleistungsstudien zur Ermittlung von Profilstrukturen genutzt. In PISA erfolgten LPA, um die Schulsysteme in den beteiligten Ländern nach ihren Beurteilungs- und Rechenschaftslegungspolitiken zu klassifizieren (OECD, 2011, S. 153 ff.). In IGLU bzw. TIMMS[31] werden diese Verfahren verwendet, um Schülerinnen und Schüler in den fächerübergreifenden Skalen und Subskalen zu gruppieren und verschiedene Leistungs- oder Kompetenzprofile zu identifizieren (Bos et al., 2012a, S. 239 ff.; Bos et al., 2012b, S. 281 ff.). Diese Ergebnisse aus IGLU/TIMSS über eine Stichprobe von 3.928 Schülerinnen und Schülern sollen im Folgenden kurz dargestellt werden, da sie als Vergleichsrahmen für die in Abschnitt 4.4.2 beschriebenen Analysen herangezogen werden.

[31] Im Erhebungszyklus 2011 wurden IGLU und TIMSS (*Trends in International Mathematics and Science Study*) zusammen durchgeführt und es wurde eine gemeinsame Stichprobe realisiert (Tarelli, Valtin, Bos, Bremerich-Vos & Schwippert, 2012, S. 19). Einige Inhalte sind in den beiden Studien daher identisch, wie auch die der LPA.

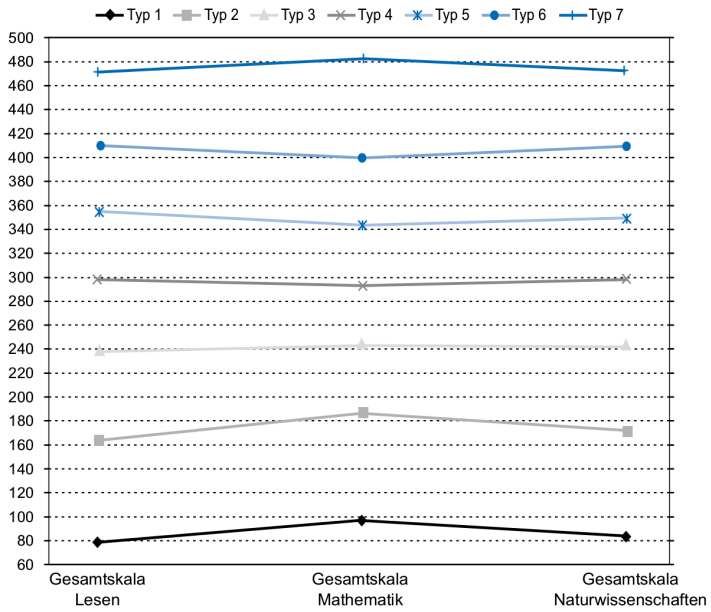



Abbildung 3.12: Leistungsprofile in IGLU/TIMSS, bezogen auf drei Kompetenzbereiche (Bos et al., 2012b, S. 283)

Die LPA-Modellwahl basierte in IGLU/TIMSS auf den informationstheoretischen Maßen. Diese wiesen für eine Modelllösung mit sieben Profilen die besten Werte aus. Für die Gesamtskalen der Kompetenzdomänen Lesen, Mathematik und Naturwissenschaften (drei Indikatorvariablen) als auch für die inhaltsbezogenen Subdomänen (acht Indikatorvariablen[32]) zeigten sich insgesamt parallele Leistungsprofile, die sich hauptsächlich durch den Niveaugrad voneinander unterschieden (Bos et al., 2012a, S. 241, 243).

Abbildung 3.12 illustriert die Leistungsprofile im Hinblick auf die Gesamtskalen. Der Mittelwert wurde auf 300 und die Standardabweichung auf einen Wert von 100 normiert. Hier zeigt sich domänenübergreifend, dass in Typ 7 mit durchschnittlich 475 Punkten die leistungsstärksten Kinder vorhanden sind. Die Leistungsmittelwerte sinken daraufhin kontinuierlich bis hin zum leistungsschwächsten Typ 1 ab, der im Durchschnitt 86 Punkte erreicht. Der Unterschied der Leistungsmittelwerte beträgt zwischen den Typen durchschnittlich 65 Punkte. Innerhalb der Profile sind sie geringer und gehen 4 bis 22 Punkte auseinander (Bos et al., 2012a, S. 240 ff.).

Bei den inhaltsbezogenen Subdomänen der drei Kompetenzbereiche sind die Profilverläufe variationsreicher (vgl. Abbildung 3.13). Am stärksten fallen die Leistungsdifferenzen bei den leistungsschwächsten und -stärksten Kindern (Typ 1 bzw. Typ 7) aus. Hier gibt es etwa 30 Punkte relative Leistungsdifferenz in Profil 1 innerhalb der naturwissenschaftlichen Subdomänen sowie 40 Punkte in Profil 7 innerhalb der mathematischen Subdomänen. Im Allgemeinen bilden größere Unterschiede innerhalb der Typen in den Subdomänen eines

[32] Lesen: *Literarisch* und *Informierend*, Mathematik: *Arithmetik*, *Geometrie/Messen* und *Umgang mit Daten* sowie Naturwissenschaften: *Biologie*, *Physik/Chemie* und *Geographie*

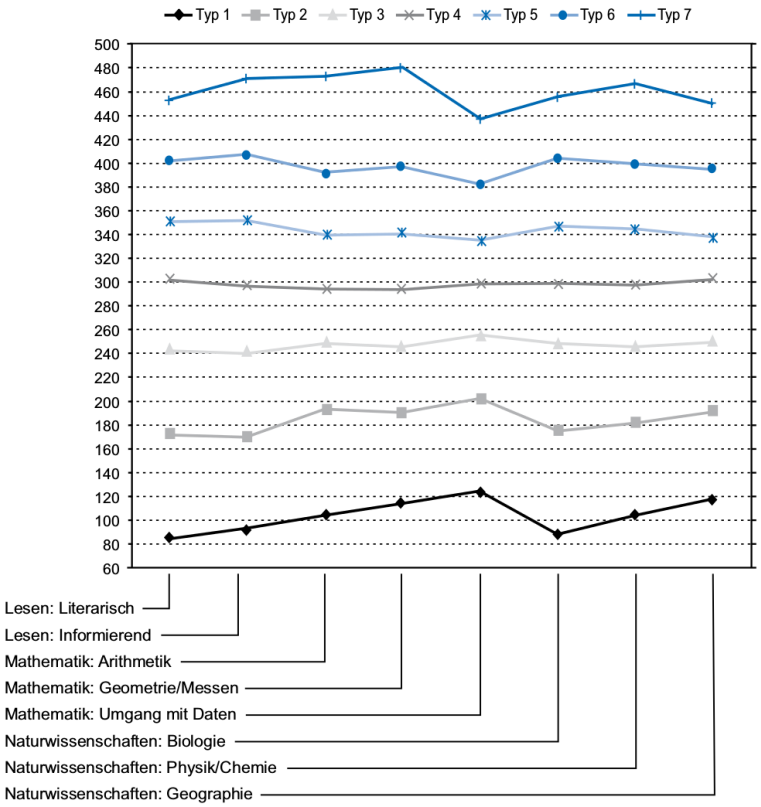



Abbildung 3.13: Leistungsprofile in IGLU/TIMSS, bezogen auf die Subdomänen der Kompetenzbereiche (Bos et al., 2012b, S. 285)

Kompetenzbereichs allerdings die Ausnahme. An einigen Stellen sind noch rund 20 Punkte Differenz abzulesen. Diese finden sich in Typ 7 in den Subdomänen des Lesens, in Typ 1 und 6 in Mathematik sowie in Typ 1, 2 und 7 in Naturwissenschaften. Bei den Kindern mit eher durchschnittlichen Leistungen, Typ 3 bis 5, sind die Differenzen besonders variationsarm. Bos et al. (2012a, S. 253) halten zusammenfassend fest: „Die Leistungen in den drei Domänen sind bei der Mehrheit der Kinder, wenn auch auf unterschiedlichen Niveaus, recht ausgeglichen. Große Unterschiede in den Leistungen sind die Ausnahme – am deutlichsten treten sie bei Schülerinnen und Schülern mit geringen und sehr geringen Leistungen auf, bei denen die Leseleistungen am schlechtesten ausfallen.“

# 4 Analyse der Daten

In diesem Kapitel werden die Analyseergebnisse der IRT-Skalierungen beschrieben. Grundlage für die Datenanalysen sind die in Abschnitt 1.2 formulierten Forschungsfragen und die in Abschnitt 3.3 beschriebenen Forschungsmethoden. Den Auftakt bildet Abschnitt 4.1, in dem zunächst auf die Auswertung ganzer Wörter eingegangen wird. Abschnitt 4.2 bildet den Schwerpunkt des Kapitels, indem kompetenzmodelldifferenzierte Befunde dokumentiert werden. Diese Befunde beinhalten eine Konstruktvalidierung (Abschnitte 4.2.1 und 4.2.2) und einen testübergreifenden Vergleich (Abschnitt 4.2.3) von gutschrift und dem SRT. Dafür werden jeweils psychometrische Kennwerte herangezogen und direkt gegenübergestellt. Zum Abschluss von Abschnitt 4.2 werden die Ergebnisse des SRT aus ZuRecht in der IGLU-Hauptuntersuchung verortet (Abschnitt 4.2.4), um diese im Kontext der großen Leistungsstudie zu überprüfen. In Abschnitt 4.3 werden die orthografischen Testleistungen auf Ganzwort- und Teilkompetenzebene mit Hintergrundmerkmalen der Schülerinnen und Schüler in Beziehung gesetzt. Abschnitt 4.4 bildet den Abschluss dieses Kapitels. Hier erfolgen Einzel- und latente Profilanalysen, wobei zunächst eine Beschreibung ausgewählter Fallbeispiele erfolgt und abschließend mit Hilfe probabilistischer Clusterverfahren Profilstrukturen ermittelt werden.

## 4.1 Globale Auswertung zum Kompetenzstand

Ein erster Eindruck über die Testschwierigkeit wird durch die Häufigkeitsverteilungen auf Ganzwortebene vermittelt, die in Tabelle 4.1 zusammengetragen sind.[1] Von den 35 Wörtern im gutschrift-Test werden durchschnittlich 20 richtig geschrieben. Das schlechteste Ergebnis mit 33 Falschschreibungen hat ein Kind und das mit 34 Richtigschreibungen beste Ergebnis erreichen sechs Schülerinnen und Schüler. Kein Kind hat den gutschrift-Test damit fehlerfrei verschriftet.

Für den SRT sind in Tabelle 4.1 Ergebnisse auf Basis des vollständigen Diktats mit 121 Wörtern und auf Basis einer selektierten Variante mit 78 Wörtern abgetragen. Da der SRT

| Test | n | Wörter-anzahl | M | Prozent korrekt | Min. | Max. | SD | VarK |
|---|---|---|---|---|---|---|---|---|
| gutschrift | 566 | 35 | 20,30 | 58% | 2 | 34 | 7,13 | 35% |
| SRT | 562 | 121 | 90,36 | 75% | 30 | 120 | 15,81 | 18% |
| SRT selektiert | | 78 | 50,17 | 64% | 5 | 77 | 14,22 | 28% |

Tabelle 4.1: Univariate Beschreibungen auf Ganzwortebene

ein Fließtextdiktat ist, beinhaltet er viele „kleine Wörter“ (größtenteils Artikel, Konjunktionen, Präpositionen und Pronomen), die keine Struktureinheiten in den Teilkompetenzen ausbilden. Zudem schreibt die Mehrheit der Kinder diese fehlerfrei. Sie bieten daher wenige Informationen über den Leistungsstand der Schülerinnen und Schüler, weshalb sie aus der 78-Wörtervariante ausgeschlossen worden sind.[2] Sie eignet sich damit besser für einen Vergleich mit dem gutschrift-Test, der hauptsächlich aus längeren, z. B. durch Kompositabildung entstandenen, Wörtern besteht (vgl. Abschnitt 3.1).

Der Mittelwert richtig geschriebener Wörter liegt im SRT bei der Schreibung des vollständigen Diktates bei 90,36. Er fällt bei der 78-Wörtervariante erwartungskonform deutlich geringer aus und liegt bei 50,17. Ebenfalls fällt die Differenz zwischen den beiden SRT-Varianten beim Minimum an korrekt geschriebenen Wörtern auf. Das leistungsschwächste Kind hat in der gekürzten Diktatform 5 Richtigschreibungen und in der vollständigen Diktatform sechsmal so viele. Bei der 121-Wörtervariante erhält die Schülerin bzw. der Schüler damit immerhin eine Rückmeldung über 1/4 korrekter Verschriftlichungen. Auch im SRT liegt kein fehlerfreies Diktat vor. Das leistungsstärkste Ergebnis erreicht in beiden Varianten dasselbe Kind mit jeweils einem Fehler.

Der *Prozent korrekt*-Wert in Tabelle 4.1 eignet sich für einen direkten Vergleich der beiden Rechtschreibtests im Hinblick auf die durchschnittliche Anzahl richtig geschriebener Wörter. Er errechnet sich aus dem Verhältnis von Mittelwert und Wörteranzahl und gibt damit den prozentualen Anteil an Richtigschreibungen an. In gutschrift wurden im Mittel 58 Prozent der Wörter normgerecht verschriftet. Im SRT betragen die Prozent-korrekt-Werte 75 Prozent und 64 Prozent (selektierte Variante). Tabelle 4.1 zeigt einen höheren Schwierigkeitsgrad des gutschrift-Tests, der im Vergleich zum SRT vermutlich über die Mehrheit an langen und komplexen Wörtern erklärt werden kann. Die Standardabweichung beträgt bei gutschrift 7,13 und bei den SRT-Varianten 15,81 bzw. 14,22. Zusätzlich wird der *Variationskoeffizient* (VarK) angegeben. Er ist definiert als:

$$\mathrm{VarK} = \frac{\sigma}{\overline{X}} \cdot 100$$

(in Anlehnung an Ghanbari, 2002, S. 150)

[1] Ein Teil der hier im Abschnitt berichteten Ergebnisse wurde bereits von der Autorin publiziert (Kowalski & Voss, 2009).

[2] Es handelt sich dabei um die folgenden 43 ausgeschlossenen Wörter: da, den, du, eine, einer, hat, in, man, nach, vor, zu; 2-mal am, auf, dem, der, es, und; 3-mal das, sie; 4-mal die, mit; 6-mal ein. Die Groß- und Kleinschreibung wird bei dieser Aufzählung nicht berücksichtigt.

| Test | leistungs-schwächste 15% | leistungs-stärkste 15% | Schreib-varianten | variantenreichstes Wort, Schreibvarianten | schwierigstes Wort, Richtigschreibungen | einfachstes Wort, Richtigschreibungen |
|---|---|---|---|---|---|---|
| gutschrift | ≤ 34% richtig | ≥ 80% richtig | 2.264 | empfindlich,127 | Schnellste, 16% | viele, 94% |
| SRT | ≤ 60% richtig | ≥ 87% richtig | 3.831 | Schnurrbarthaaren, 163 | Schnurrbarthaaren, 11% | du, 100% |
| SRT selektiert | ≤ 42% richtig | ≥ 82% richtig | 3.529 | | | fragt, 95% |

Tabelle 4.2: Perzentile, Schreibvarianten und Wortschwierigkeiten

Die Standardabweichung $\sigma$ wird hier auf den Mittelwert $\overline{X}$ bezogen und als Prozentangabe wiedergegeben. Es handelt sich um ein Maß der relativen Streuung, mit dem die Variation der Werte unabhängig von unterschiedlichen Mittelwerten miteinander verglichen und beurteilt werden kann (Kohn & Öztürk, 2013, S. 70). Beim SRT weichen die Leistungswerte im Durchschnitt weniger vom Mittelwert ab als im gutschrift-Test (VarK = 35 Prozent): Sie variieren um 18 bzw. 28 Prozent um den Mittelwert und weisen damit eine höhere Homogenität auf (vgl. Tabelle 4.1).

Für den gutschrift-Test ist ein direkter Vergleich des Abschneidens der Kinder aus ZuRecht mit den Stichproben aus den repräsentativen IGLU-Ergänzungsstudien 2001 und 2006 über 15 Ankerwörter möglich. Hierbei handelt es sich um identische Testwörter, die sowohl in ZuRecht als auch in den beiden IGLU-Erhebungen eingesetzt wurden.[3] Der Mittelwert an Richtigschreibungen liegt in IGLU 2001 bei 51 Prozent und in IGLU 2006 bei 58 Prozent. In ZuRecht beträgt er 52 Prozent, womit er leicht oberhalb des Wertes der ersten IGLU-Erhebung ist und bezogen auf die zweite IGLU-Erhebung 6 Prozentpunkte schlechter ausfällt. Ebenfalls sollen die Resultate des SRT mit denen aus der IGLU-Voruntersuchung in Beziehung gesetzt werden. Die Ergebnisse im Pretest betreffen eine SRT-Variante mit 103 Wörtern, die mit dem Diktat in ZuRecht eine Schnittmenge von 70 identisch eingesetzten Testwörtern aufweist. Der errechnete Prozent-korrekt-Wert liegt für diese gemeinsamen Testwörter in der Voruntersuchung bei 84 Prozent und in ZuRecht bei 76 Prozent. Damit schneiden die Schülerinnen und Schüler aus dem IGLU-Pretest 8 Prozentpunkte besser ab als die Kinder aus ZuRecht.

In Tabelle 4.2 wird das 15. und 85. Perzentil der Stichprobe betrachtet, um die Werte der leistungsschwächsten und -stärksten Kinder gegenüberzustellen. Die schlechtesten 15 Prozent der Kinder schreiben bei gutschrift maximal 12 der 35 Testwörter (34 Prozent) richtig. Beim SRT werden in dieser Gruppe höchstens 73 bzw. 33 (selektierte Variante) Wörter richtig geschrieben. Der Anteil an Richtigschreibungen von maximal 60 bzw. 42 Prozent liegt damit in beiden SRT-Testvarianten über dem von gutschrift. In der Gruppe der oberen 15 Prozent kann dies analog beobachtet werden: Der Anteil an Richtigschreibungen beträgt bei gutschrift mindestens 80 Prozent (28 und mehr Wörter fehlerfrei) und beim SRT mindestens 87 bzw. 82 Prozent (105 bzw. 64 und mehr Wörter fehlerfrei). Die Diskrepanz zwischen den Gruppen fallen hoch aus. Der Abstand ist mit mindestens 46 Prozentpunkten im gutschrift-Test am größten; beim SRT beträgt er 27 bzw. 40 Prozentpunkte. In dem

[3] Eine vergleichende Auswertung in Abhängigkeit von unterschiedlichen Hintergrundmerkmalen über die Ankerwörter für die IGLU-Erhebungszeitpunkte 2001 und 2006 findet sich in Kowalski et al. (2010).

| gutschrift | |
|---|---|
| *entfindlich | 54 |
| *emfindlich | 33 |
| *endfindlich | 27 |
| *empfintlich | 26 |
| *entfintlich | 25 |
| *emfintlich | 15 |
| *enfindlich | 8 |
| *endpfindlich | 7 |
| *enpfindlich | 7 |
| *infindlich | 6 |

| SRT | |
|---|---|
| *Schnurbarthaaren | 118 |
| *Schnurbardhaaren | 27 |
| *Schnurbartharen | 23 |
| *Schnurbartharren | 15 |
| *schnurbarthaaren | 14 |
| *Schnurbart Haaren | 12 |
| *Schnurbadhaaren | 9 |
| *Schnurbart haaren | 9 |
| *Schnurbathaaren | 9 |
| *Schnurrbardhaaren | 9 |

Tabelle 4.3: Die häufigsten Falschschreibungen der variantenreichsten Wörter

reduzierten SRT werden von den leistungsschwächsten 15 Prozent der Kinder höchstens ca. halb so viele Wörter richtig verschriftet, die mindestens von der Gruppe der oberen 15 Prozent normgerecht geschrieben werden. Bei gutschrift sind dies etwa 43 Prozent. Dieser Befund zeigt den großen Leistungsunterschied zwischen den rechtschreibschwächeren und -stärkeren Schülerinnen und Schülern.

Tabelle 4.2 beinhaltet ebenfalls Angaben zu den unterschiedlichen Schreibvarianten. Insgesamt wurden die 35 Wörter des gutschrift-Tests in 2.264 verschiedenen Varianten verschriftet. Das variantenreichste Wort ist empfindlich mit 127 unterschiedlichen Schreibweisen (inklusive der richtigen). Die zehn häufigsten Varianten von Falschschreibungen sind in Tabelle 4.3 aufgelistet. Schwierigkeiten bereiten bei dem Wort empfindlich danach insbesondere die lautanalytische Verschriftung von ⟨m⟩, die Verschriftung der Affrikate in Form von ⟨pf⟩ sowie die Auslautverhärtung von ⟨d⟩. Aus den Schreibvarianten wird deutlich, dass die Schülerinnen und Schüler teilweise die Bedeutung des Wortes nicht erkannt haben. Wenige Probleme verursacht hingegen das Suffix ⟨lich⟩.

Die Gesamtanzahl der Schreibvarianten im SRT beträgt 3.831 bzw. 3.529 (selektierte Variante) (vgl. Tabelle 4.2). Theoretisch kämen damit im Mittel in der langen SRT-Version 32 Schreibweisen auf ein Testwort und in der gekürzten SRT-Version 45 Schreibweisen, während es in gutschrift 65 sind. Im SRT bildet Schnurrbarthaaren das variantenreichste und zugleich schwierigste Wort. Es wurde 499-mal falsch (89 Prozent) und 162-mal unterschiedlich falsch geschrieben. Auffällig ist, dass die häufigste Falschschreibung (*Schnurbarthaaren mit 21 Prozent bzw. 118-mal) fast doppelt so oft auftritt wie die Richtigschreibung (11 Prozent bzw. 63-mal). Sie ergibt sich aus dem Fehlen des doppelten Konsonatengraphems ⟨rr⟩, wie Tabelle 4.3 zeigt, und könnte z. B. mit der Assoziation von „die Schnur“ anstatt „das Schnurren“ erklärt werden. Daneben bereiten das ⟨t⟩ sowie die Vokalgraphemverdopplung ⟨aa⟩ Probleme. Die Schreibung des Wortanfangs und -endes (⟨sch⟩ für /ʃ/ sowie des Suffixes ⟨en⟩) erfolgt zumeist sicher.

Das schwierigste Wort in gutschrift ist Schnellste. Der Anteil an richtigen Schreibungen beträgt 16 Prozent (92-mal). Die Falschschreibungen sind größtenteils auf Fehler bei der Groß- und Kleinschreibung zurückzuführen. Die Substantivierung wurde, bei ansonsten

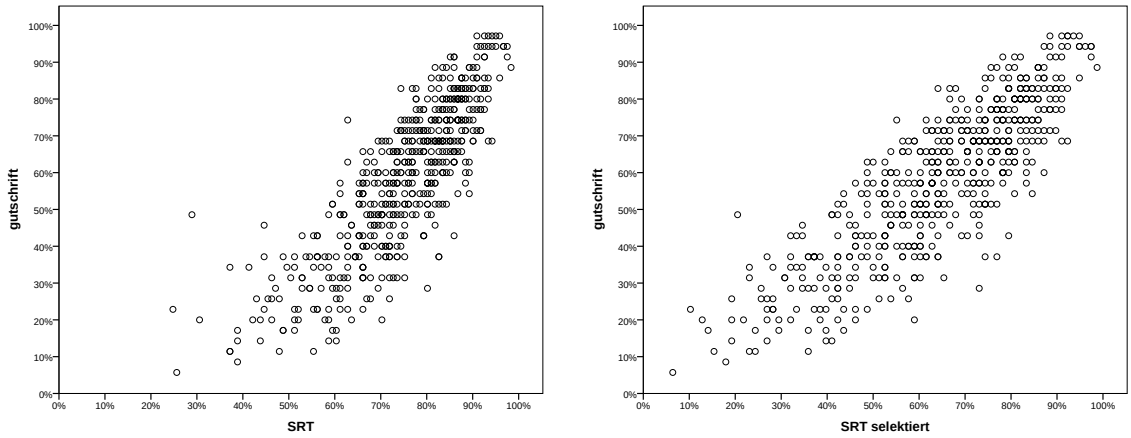



Abbildung 4.1: Streudiagramme der durchschnittlichen prozentualen Testleistung

richtiger Schreibung (*schnellste), 299-mal (53 Prozent) nicht erkannt. Das heißt, dass der Gesamtfehleranteil des Wortes (84 Prozent, 474-mal) in fast zwei Drittel der Fälle (63 Prozent) aus der fälschlichen Kleinschreibung resultiert. Das einfachste Wort in gutschrift lautet viele und wurde zu 94 Prozent (529-mal) korrekt verschriftet (drei Kinder haben das Testwort ausgelassen und wurden als Missings kodiert). Beim SRT ist du mit 560 (gerundet 100 Prozent) Richtigschreibungen, bzw. in der selektierten Variante fragt mit 536 (95 Prozent) Richtigschreibungen, das leichteste Wort.

### Korrelationen zwischen den Tests

In Abbildung 4.1 ist die gemeinsame Verteilung der durchschnittlichen Leistung für die beiden Tests auf Basis der Prozent-korrekt-Werte dargestellt. Verwendet wurden hierfür die Testergebnisse der gemeinsamen Stichprobe von 547 Schülerinnen und Schülern. Ein deutliches Muster zeichnet sich bei der Verteilung der Variablen ab; die Punktewolke folgt einem linearen Trend. Erwartungsgemäß erzielen Schriftlernende bei der Schreibung ganzer Wörter in beiden Tests ähnliche Ergebnisse. Der positive Zusammenhang weist einen Korrelationskoeffizienten (Korrelation nach Pearson) von $r = 0,83$ (SRT) und $r = 0,85$ (SRT selektiert) auf. Die Punktewolken sind nach rechts verschoben, da der SRT leichter ist, wobei diese Beobachtung bei der nicht selektierten Testversion stärker zutrifft.

### Zusammenschau

In diesem Abschnitt konnte ein erster Überblick über den Leistungsstand der Schülerinnen und Schüler durch die Ergebnisdarstellungen auf Ganzwortebene gewonnen werden. Auffällig ist die hohe Anzahl an Schreibvarianten der Schülerinnen und Schüler sowie die Höhe der Differenz zwischen leistungsstarken und -schwachen Schreibern. Dies deutet zum einen auf eine Verunsicherung der Kinder bezüglich der Orthografie, zum anderen auf eine Heterogenität des Leistungsstandes in der Untersuchungsgruppe hin. Die Befunde

sind damit konform zu den Ergebnissen aus z. B. KESS 4 oder IGLU 2001 (vgl. Abschnitte 2.1.1 und 2.2.2). Auf der Ebene ganzer Wörter konnte eine höhere Schwierigkeit und Leistungsvariation des gutschrift-Tests im Vergleich zum SRT festgestellt werden. Der Leistungsstand der ZuRecht-Stichprobe wurde zudem im Rahmen des gutschrift-Tests mit Referenzwerten aus IGLU-E 2001 in Beziehung gesetzt. Demnach haben die Kinder aus ZuRecht einen ähnlichen Leistungsstand wie die Kinder des Bundesdurchschnitts. Für weiterführende Vergleiche zwischen gutschrift und dem SRT sind insbesondere die Ergebnisse auf Kompetenzmodellebene interessant. Im folgenden Abschnitt 4.2 werden dafür analog zu Tabelle 4.1 die Häufigkeitsverteilungen für die differenzierten Teilkompetenzen der beiden Tests dargestellt.

## 4.2 Differenzielle Auswertung und Validierung der Kompetenzmodelle

In diesem Abschnitt werden die beiden Rechtschreibtests gutschrift-diagnose und SRT kompetenzmodellbasiert ausgewertet, validiert und empirisch miteinander verglichen. In den Abschnitten 4.2.1 und 4.2.2 erfolgt die Darstellung der Analyseergebnisse zunächst getrennt für die Tests. In diesen beiden Abschnitten wird als erstes der Itemreviewprozess beschrieben, indem die Anzahl modellkonformer Aufgaben dokumentiert und anhand der Wright Map das Verhältnis von Itemschwierigkeit und Personenfähigkeit veranschaulicht wird. Als zweites wird auf bivariater Ebene die Struktur der Daten durch latente Interkorrelationen quantifiziert. Auf Gesamtmodellebene erfolgen als drittes Vergleiche von alternativen Kompetenzmodellen via Modellgeltungstests. Bei gutschrift wird in diesem Zusammenhang zunächst ein Modell mit ordinalem Skalenniveau (Partial-Credit-Modell) dem Raschmodell gegenübergestellt, um das Modell mit der besten Datenanpassung auszuwählen und für die weiterführenden Analysen zu nutzen. In Abschnitt 4.2.3 werden die beiden Rechtschreibtests im direkten Vergleich betrachtet, indem die Ergebnisse einer gemeinsamen, testübergreifenden Skalierung beschrieben werden. Hier werden differenzierte Befunde zu den Teilkompetenzen gegenübergestellt, worüber ein Einblick in die Schwierigkeiten der Subskalen und den Kompetenzstand der Schülerinnen und Schüler gegeben wird. Im Anschluss erfolgt eine Korrelations- und Reliabilitätsanalyse über die Teilkompetenzen der beiden Tests. Die Werte dieser Statistiken können durch das neundimensionale Modell der gemeinsamen Skalierung unmittelbar miteinander in Beziehung gesetzt werden. Abgeschlossen wird der Abschnitt durch eine Einbindung der IGLU-Hauptuntersuchung (Abschnitt 4.2.4). Hierüber soll der Kompetenzstand der in ZuRecht getesteten Kinder eingeordnet werden und eine Prüfung der Befunde aus ZuRecht erfolgen.

### 4.2.1 gutschrift-diagnose

#### Datenaufbereitung

Die Daten aus dem gutschrift-Expertenprogramm (vgl. Abschnitt 3.2) mussten zunächst für die Skalierung aufbereitet werden. Sie lagen in ordinaler Datenstruktur vor, sodass es innerhalb eines Wortes mehrere mögliche Fehlerquellen[4] bzw. Indikatoren für eine Teilkompetenz gab und unterschiedlich viele Punktezahlen bei der Schreibung erreicht werden konnten. Da es sich um eine fehlerbasierte Auswertung handelt (vgl. Abschnitt 2.2), wurden die Daten für die Raschskalierung numerisch aufsteigend von falsch nach richtig rekodiert: 0 ≙ „null Indikatoren richtig geschrieben“, 1 ≙ „ein Indikator richtig geschrieben“, 2 ≙ „zwei Indikatoren richtig geschrieben“ etc. In einem weiteren Schritt mussten die in den Daten enthaltenen Informationen verdichtet[5] werden: Bei 14 Fehlerquellen fielen „Lücken“ in den Antworten bei den Häufigkeitsauszählungen auf, die geschlossen werden mussten, damit jede Abstufung Datenmaterial enthält. Hierfür wurde jeweils die nächsthöhere Antwortkategorie um einen Wert verringert. Wenn also z. B. bei einem 3-stufigen Item kein Kind genau einen Indikator richtig geschrieben hat und damit eine Lücke zwischen den Antwortkategorien 0 und 2 entstanden ist, so wurde der Wert 2 in 1 rekodiert. Um eine stabile Anzahl an Schwellen zu besitzen, erfolgte zudem eine Deckelung der Fehlerquellen innerhalb eines Wortes einer Teilkompetenz, sodass im Endergebnis vier mögliche Antwortabstufungen in den Daten vorlagen (0, 1, 2, 3).

Im Anschluss an die Datenaufbereitung wurden die Daten mit dem Partial-Credit-Modell (PCM) skaliert. Das PCM ist ein einparametriges Modell für die Auswertung mehrstufiger Daten, bei dem richtige, teilweise richtige und falsche Antworten/Lösungen voneinander unterschieden werden (Rost, 2004, S. 209). Es werden also nicht ausschließlich, wie beim Raschmodell (RM), „falsch“ und „richtig“ unterschieden, sondern mehrere Antwortkategorien berücksichtigt, wobei deren Anzahl variieren kann.[6] Dementsprechend wird die Wahrscheinlichkeit modelliert, eine bestimmte Antwortkategorie mit einem hohen Punktewert zu erreichen, und nicht die Wahrscheinlichkeit einer generell richtigen Antwort. Im Folgenden soll das PCM nicht näher ausgeführt werden, da es – wie sich zeigen wird – für die weitergehenden Analysen dieser Arbeit keine Relevanz hat. Stattdessen sei auf nähere Informationen z. B. in Müller (1999) verwiesen.

#### Modellvergleich: RM und PCM mit vier- und zweidimensionaler Struktur

Die gutschrift-Daten wurden zunächst mit dem PCM ausgewertet ($n = 566$), bei dem mehrstufige Items berücksichtigt werden, um möglichst vollständig alle Informationen, also wie viele Fehler bei einem Wort innerhalb einer Teilkompetenz gemacht wurden, zu erhalten. Neben diesem vierdimensionalen Modell (4D PCM) wurde auch ein zweidimensionales

[4] vom Testautorinnenteam gewählte Begrifflichkeit
[5] vgl. Wu et al. (2007, S. 99)
[6] Das PCM wird daher auch als ordinales Raschmodell bezeichnet (Rost, 2004, S. 209).

| Modell | Deviance | Parameter | AIC | BIC | CAIC | verglichene Modelle | Δ Deviance | df |
|---|---|---|---|---|---|---|---|---|
| 4D PCM | 45.657,45 | 267 | 46.191,45 | 47.349,85 | 47.616,85 | | | |
| 4D RM | 44.674,73 | 140 | 44.954,73 | 45.562,14 | 45.702,14 | 4D PCM vs. 4D RM | -982,72 | 127 |
| 2D PCM | 45.688,97 | 260 | 46.208,97 | 47.337,00 | 47.597,00 | | | |
| 2D RM | 44.742,78 | 133 | 45.008,78 | 45.585,81 | 45.718,81 | 4D RM vs. 2D RM | 68,05 | 7 |

Tabelle 4.4: gutschrift-Modellvergleich von PCM und RM

„Fähigkeiten“-Modell (2D PCM) berechnet. Bei dem zweidimensionalen Modell wurde über die zwei phonographischen Fähigkeiten (elementar und erweitert) sowie über die zwei grammatischen Fähigkeiten jeweils eine Dimension gebildet. Es wurde modelliert, da mit den Begrifflichkeiten „elementar“ und „erweitert“ Schwierigkeitsgrade (oder auch Niveaus oder Stufen) beschrieben werden, aber aus psychometrischer Sicht damit nicht zwingend Dimensionen mit unterschiedlichen latenten Personenfähigkeiten zu verstehen sind. Die Unterscheidung des gutschrift-Kompetenzmodells in vier Teilkompetenzen sollte daher validiert werden.

Im Zuge der Skalierung beider Modelle (4D PCM und 2D PCM) zeigten sich viele auffällige Items mit modellabweichenden Werten. Aus diesem Grund wurden zwei alternative Modelle berechnet: ein vier- sowie ein zweidimensionales Raschmodell (4D RM und 2D RM) als entsprechende Pendants. Bei einem ersten Modelldurchlauf der Raschmodelle entsprachen, im Vergleich zu den PCMs, deutlich mehr Aufgaben den Kriterien für ein „gutes“ Item (vgl. Abschnitt 3.3.2.2). Neben der Itemanalyse verdeutlichten auch die informationstheoretischen Kriterien eine bessere Modellanpassung der dichotomen Modelle, wie aus Tabelle 4.4 abzulesen ist. Die Werte aller Informationsindizes sowie der Likelihoodquotiententest fallen für die Raschmodelle niedriger aus, und innerhalb dieses Skalierungsmodells für das vierdimensionale Modell am niedrigsten.

Inferenzstatistische Vergleiche mittels $\chi^2$-verteilter Testprüfgröße sichern diesen Befund ab. Bei dem Vergleich der beiden vierdimensionalen Modelle ist der errechnete Wert der Prüfgröße von -982,72 kleiner als der kritische Wert der $\chi^2$-Verteilung mit 127 Freiheitsgraden bei einem Signifikanzniveau von $\alpha = 0,01$ (vgl. Tabelle 4.4). Damit erklärt das PCM die Daten nicht signifikant besser als das RM. Das RM kann damit als ein mit den Daten verträgliches Modell betrachtet werden. Ein Vergleich der beiden Raschmodelle spricht ebenfalls für eine bessere Anpassungsgüte des 4D RM gegenüber dem 2D RM, da die Prüfgröße größer als der kritische Wert ist und damit die Wahrscheinlichkeit für die Geltung des restriktiveren Modells unter 1 Prozent liegt.

Aus den Analyseergebnissen von Tabelle 4.4 lässt sich folgern, dass die gutschrift-Daten unter Anwendung des Raschmodells skaliert werden können und eine bessere Datenanpassung als die PCMs aufweisen. Das RM sollte daher, unter Rückgriff auf das theoretische Postulat des Einfachheitskriteriums (vgl. Abschnitt 3.3.2.1), für die weiterführenden Auswertungen genutzt werden. Innerhalb der Raschmodelle erweist sich die theoretisch angenommene vierdimensionale Kompetenzstruktur vorteilhafter als die zweidimensionale

Struktur. Das vierdimensionale Raschmodell bildet daher für alle weiteren Datenauswertungen von gutschrift die Basis und wird im Folgenden nur kurz als 4D-Modell bezeichnet, da es seine Favorisierung gegenüber den PCM-Modellen und dem zweidimensionalen Raschmodell bewiesen hat.

### Itemreview

Wie in Abschnitt 3.3.2 beschrieben, lassen sich die Aufgabenschwierigkeiten und Personenfähigkeiten auf einer gemeinsamem Skala abbilden, auch wenn sie getrennt voneinander geschätzt werden. Dies ist eine für die Fachwissenschaft nützliche Eigenschaft probabilistischer Testverfahren, da durch die Quantifizierung der Itemschwierigkeiten Aufschlüsse über die strukturellen Bestandteile, die zu dem Schwierigkeitsgrad geführt haben, gewonnen werden können. Bei Klieme, Avenarius et al. (2007, S. 74 ff.) werden hierüber Differenzierungen des Kompetenzniveaus (Stufenmodelle) vorgenommen.

Abbildung 4.2 zeigt die Gegenüberstellung der Aufgabenschwierigkeiten und Personenfähigkeiten anhand der Wright Map für den vierdimensionalen Skalierungslauf des Raschmodells mit den Teilkompetenzen bzw. Dimensionen:

1. Elementar phonographisch
2. Elementar grammatisch
3. Erweitert phonographisch
4. Erweitert grammatisch

Grundlage für die Skalierung bilden die 82 modellkonformen Items, die im Rahmen der Itemanalyse (vgl. Abschnitt 3.3.2.2) in dem Modell verblieben sind. Insgesamt wurden 48 Aufgaben mit unzureichendem Itemfit (niedriger als 0,8 oder höher als 1,2) und/oder geringer Trennschärfe (kleiner als 0,2) eliminiert. Am linken Abbildungsrand in Abbildung 4.2 ist die Skalierungsmetrik in Logiteinheiten dargestellt, auf der die geschätzten Personen- und Itemparameter abgetragen sind. Die Wright Map stellt die Fähigkeits- und Schwierigkeitsmaße gegenüber (Bond & Fox, 2007, S. 292). Die Aufgabenschwierigkeitsverteilung wird durch die Zahlen 1 bis 82, die jeweils für ein Item stehen, veranschaulicht. Die Metrik der Skala wird festgelegt, indem die mittlere Aufgabenschwierigkeit auf den Wert 0 fixiert wird (Rauch & Hartig, 2008, S. 241). Die X visualisieren die Verteilung der Personenparameterschätzung, wobei jeweils fünf Kinder durch ein X repräsentiert werden. Jede Dimension hat dabei ihre eigene Spalte mit den Schätzungen der Personenfähigkeiten. Im oberen Bereich der Logitskala befinden sich die Schülerinnen und Schüler mit höherer (orthografischer) Fähigkeit sowie die Aufgaben mit höherem Schwierigkeitsgrad (Osteen, 2010, S. 77).

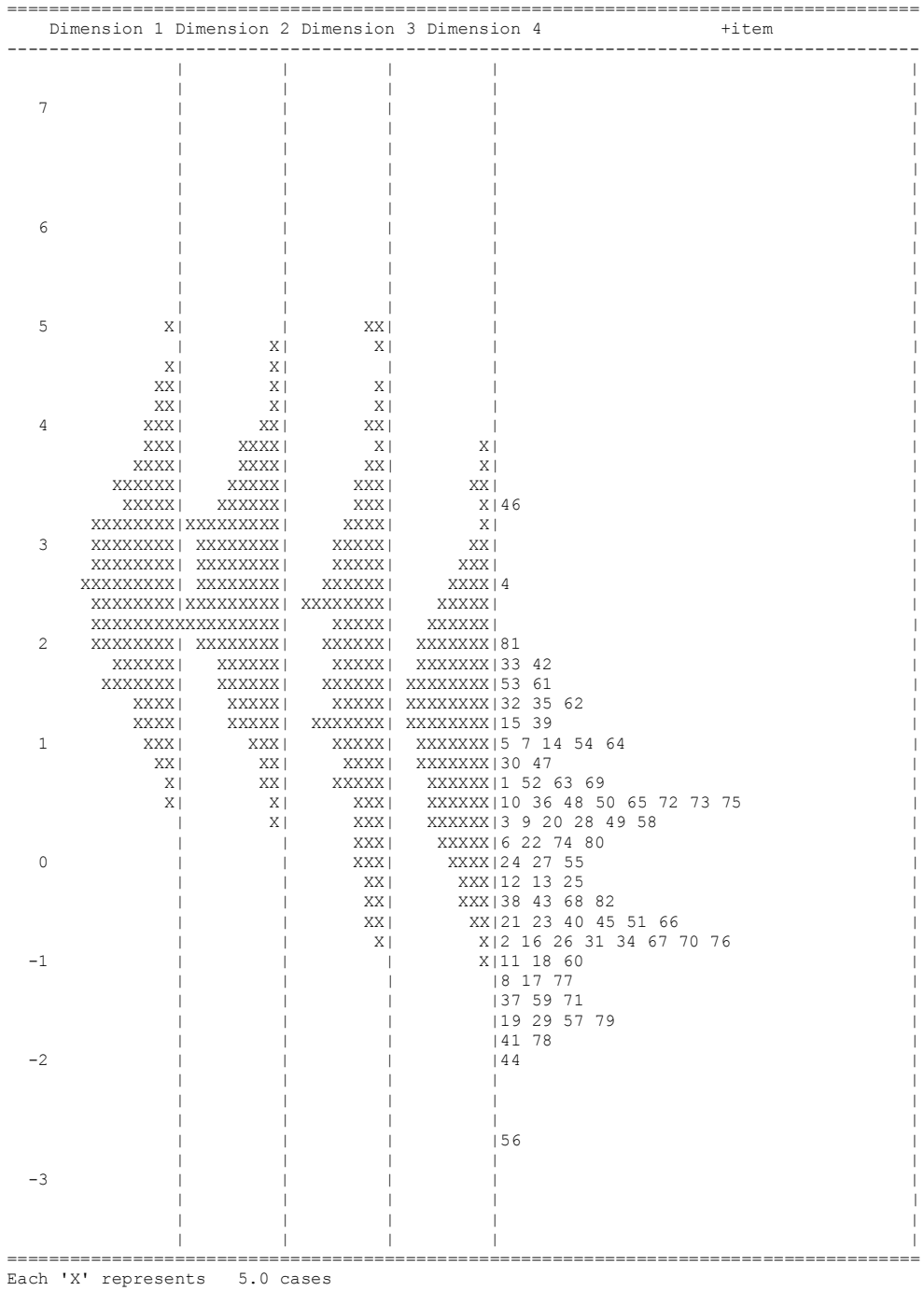



Abbildung 4.2: Wright Map für das vierdimensionale gutschrift-Modell

Die Schwierigkeitsmaße liegen im Bereich von -2.614 bis 3.329 Logits und die Fähigkeitsmaße im Bereich von -1 bis 5 Logits. Die Fähigkeiten der Schülerinnen und Schüler streuen bei den erweiterten Kompetenzdimensionen stärker als bei den elementaren. Sie nehmen Werte von 0,955 in der ersten Dimension, 0,881 in der zweiten Dimension, 1,987 in der dritten Dimension und 1,157 in der vierten Dimension an. Die Mittelwerte für die Personenparameter[7] und die Schwierigkeitsparameter liegen bei 2,557 und 0,041 (Dimension 1), 2,506 und 0,039 (Dimension 2), 1,913 und 0,059 (Dimension 3) sowie 1,285 und 0,045 (Dimension 4) Logits. Da die Personenfähigkeiten im Mittel die Aufgabenschwierigkeiten übersteigen, fallen die durchschnittlichen Lösungswahrscheinlichkeiten höher als 50 Prozent aus. Setzt man beispielsweise für die erweiterte grammatische Kompetenz die Mittelwerte der Personen- und Itemparameter in Gleichung 3.1 aus Abschnitt 3.3.2 ein, so ergibt sich, dass die Schülerinnen und Schüler für die Items dieser Dimension eine durchschnittliche Lösungswahrscheinlichkeit von 78 Prozent haben. Den Kindern der ZuRecht-Studie fallen die Aufgaben aus der erweiterten grammatischen Kompetenz und aus den weiteren Kompetenzen des gutschrift-Tests eher leichter. Dies verdeutlicht auch die linksschiefe Verteilung die in Abbildung 4.2 erkennbar ist (Bond & Fox, 2007, S. 61). Im unteren Leistungsbereich werden die Kinder gut differenziert; bei einer zukünftigen Testkonstruktion sollte der Test aber um schwierigere Items angereichert werden, um mehr Informationen zum mittleren und oberen Bereich der Fähigkeitsskala zu erfassen.

### Gegenüberstellung mit dem Generalfaktormodell

Nachdem nun im ersten Schritt das vierdimensionale Raschmodell die in den Daten enthaltenen Informationen besser abbilden konnte als die PCMs und das zweidimensionale Raschmodell sowie im zweiten Schritt die Items des Modells gereviewt worden sind, soll nun in einem dritten Schritt ein weiterer Modellvergleich erfolgen. Dabei wird dem favorisierten 4D-Modell ein *Generalfaktormodell* gegenübergestellt. Generalfaktormodelle nehmen eine übergeordnete Personenfähigkeit über alle Dimensionen hinweg an. Damit ein Vergleich möglich ist, muss sich das Modell auf denselben Datensatz und dieselben Indikatoren stützen wie das 4D-Modell. Es unterscheidet aber keine Teilkompetenzen voneinander, da es eine eindimensionale Struktur aufweist. Mit der Gegenüberstellung der beiden Modelle soll geprüft werden, ob die theoretische Ausdifferenzierung der Subskalen sinnvoll ist oder ein Modell, das Rechtschreibkompetenz als globales Konstrukt beschreibt, besser auf die Daten anzuwenden ist.

Die Ergebnisse des Vergleichs der beiden konkurrierenden Modelle sind in Tabelle 4.5 dargestellt. Allein der AIC spricht für eine bessere Anpassungsgüte des Generalfaktormodells. Die Werte des BIC und CAIC, die komplexere Modelle eigentlich wegen der stärkeren Gewichtung der Parameteranzahl mehr bestrafen, fallen hingegen für das 4D-Modell besser

[7] Die einzelnen Mittelwerte der Personenfähigkeiten können nicht unmittelbar miteinander verglichen werden, da sie keinen gemeinsamen Ursprung haben (weil die mittlere Aufgabenschwierigkeit jeder Dimension auf 0 gesetzt worden ist, s. o.). Ein Vergleich mit den Verteilungen der Itemparameter, aus denen sie ermittelt wurden, ist aber möglich (Wu et al., 2007, S. 94).

| Modell | Deviance | Parameter | AIC | BIC | CAIC | Δ Deviance | df |
|---|---|---|---|---|---|---|---|
| 4D | 35.046,07 | 92 | 35.230,07 | 35.629,22 | 35.721,22 | 146,74 | 9 |
| 1D | 35.192,81 | 83 | 35.192,81 | 35.718,91 | 35.801,91 | | |

Tabelle 4.5: gutschrift-Modellvergleich von 4D-Modell und Generalfaktormodell

aus. Eine Absicherung dieser Befunde bietet die Überführung des Likelihoodquotienten in eine $\chi^2$-verteilte Prüfstatistik, die ein signifikantes Testergebnis erzielt: Die Prüfgröße nimmt einen Wert von 146,74 an, was bei 9 Freiheitsgraden oberhalb der kritischen Grenze liegt. Die Wahrscheinlichkeit für eine Geltung von $L_0$, also des Generalfaktormodells, liegt damit unter 1 Prozent. Aus formalstatistischer Perspektive ist demnach das vierdimensionale Modell zu favorisieren.

### Korrelationsstatistische Analyse

Eine weitere wichtige Möglichkeit, um die Dimensionalität der Daten zu prüfen, kann – neben den globalen Modellprüfungstests und -vergleichen – über eine korrelationsstatistische Analyse der Teilkompetenzen erfolgen. Dabei werden die bivariaten Zusammenhänge zwischen den Teilkompetenzen quantifiziert. Sie werden daher auch als *intra-domain-Korrelationen* bezeichnet. Aus einem hohen Korrelationskoeffizienten kann abgeleitet werden, dass es nach statistischen Gesichtspunkten nicht sinnvoll ist, Teilkompetenzen zu differenzieren, da diese redundante Informationen enthalten. Über die Korrelationsanalyse wird also ebenfalls getestet, ob sich die theoretisch angenommene Struktur von Rechtschreibkompetenz anhand der empirischen Daten bestätigen lässt. Die latenten Korrelationen sind direkt als Modellparameter schätzbar, sodass sie nicht von den Messfehlern der Personenparameterschätzungen beeinflusst sind (Wu & Adams, 2006, S. 104; Rost, 2004, S. 264). Sie fallen generell höher aus als Produkt-Moment-Korrelationen und sind nicht mit diesen zu verwechseln (OECD, 2012, S. 194; Bos, Valtin, Voss, Hornberg & Lankes, 2007, S. 91).

Um die nachfolgenden Ergebnisse besser einschätzen zu können, sollen zunächst einige Korrelationswerte aus anderen Studien kurz zusammengefasst berichtet werden. Sie sollen als Referenz genutzt werden, um die Enge des Zusammenhangs zu beurteilen, also die Frage zu beantworten, wie hoch Korrelationen ausfallen dürfen, um Teilkompetenzen als relativ eigenständig bewerten zu können. Zunächst sind aus den internationalen Berichtsbänden von PISA die fächerspezifischen Ergebnisse für die im jeweiligen Erhebungszyklus fokussierten Domänen dargestellt, und im Anschluss die fächerübergreifenden.

- PISA 2000: Die dokumentierten Korrelationen zwischen den verschiedenen Leseverstehensprozessen *retrieving information*, *interpreting text* und *reflection and evaluation* liegen bei 0,97, 0,89 und 0,93 (Adams & Carstensen, 2002, S. 153).

- PISA 2003: Zwischen den mathematischen Subskalen *space and shape*, *change and relationships*, *uncertainty* und *quantity* betragen die Zusammenhänge 0,90, 0,89, 0,90, 0,92, 0,93 und 0,90 (OECD, 2005, S. 190).
- PISA 2006: Die Korrelationen der Subskalen *explaining phenomena scientifically*, *identifying scientific issues* und *using scientific evidence* aus dem naturwissenschaftlichen Bereich nehmen Werte von 0,89, 0,93 und 0,90 an (OECD, 2009, S. 215).
- PISA 2009: Die Zusammenhangszahlen lauten 0,96, 0,93 und 0,95 für die wiederholten Analysen zum Lesen mit den Subskalen *access and retrieve*, *integrate and interpret*, *reflect and evaluate* (OECD, 2012, S. 195).
- PISA 2012: Die Werte für die vier Inhaltsbereiche mathematischer Kompetenz – *quantity*, *uncertainty and data*, *space and shape* und *change and relationships* – sind 0,87, 0,84, 0,91, 0,85, 0,92 und 0,90 (OECD, 2014, S. 231).
- PISA 2000 bis 2012: Die Zusammenhänge zwischen den verschiedenen fachübergreifenden Skalen werden als *inter-domain-Korrelationen* bezeichnet. Diese fallen in PISA 2000 bis PISA 2012 folgendermaßen aus: Sie betragen zwischen Lesen und Mathematik 0,82 (im Jahr 2000), 0,77 (in 2003), 0,79 (in 2006), 0,84 (in 2009) und 0,86 (in 2012), zwischen Lesen und Naturwissenschaften (Werte in bleibender chronologischer Reihenfolge) 0,89, 0,82, 0,83, 0,87 und 0,88 sowie zwischen Mathematik und Naturwissenschaften 0,85, 0,82, 0,88, 0,89 und 0,90 (OECD, 2014, S. 230; OECD, 2012, S. 194; OECD, 2009, S. 215; OECD, 2005, S. 189; Adams & Carstensen, 2002, S. 153).

Neben PISA werden Vergleichswerte zum Lesen aus IGLU berichtet. In IGLU 2001 nehmen die latenten Korrelationen bei den differenzierten Leseverstehensprozessen *Erkennen und Wiedergeben explizit angegebener Informationen* (1), *Einfache Schlussfolgerungen ziehen* (2), *Komplexe Schlussfolgerungen ziehen und begründen; Interpretieren des Gelesenen* (3) sowie *Prüfen und Bewerten von Inhalt und Sprache* (4) die Werte 0,89, 0,88, 0,84, 0,90, 0,87 und 0,87 an (Bos et al., 2003, S. 80). Die Korrelationsanalyse der unterschiedenen zwei textsortenspezifischen Dimensionen (Leseintention) *Lesen literarischer Texte* (A) und *Lesen von Informationstexten* (B) ergab einen Zusammenhang von 0,89. (Bos et al., 2003, S. 82).

In IGLU 2006 fallen die Korrelationen der o. g. Leseverstehensaspekte (1 bis 4) höher aus und liegen bei 0,97, 0,92, 0,92, 0,94, 0,93 und 0,98 (Bos, Valtin, Voss et al., 2007, S. 91). Daraufhin wurde dazu übergegangen, nicht das vierdimensionale Modell bei der Berichtslegung zu nutzen, obwohl der CAIC, verglichen mit alternativen Modellen, die beste Modellanpassung zeigte (Bos, Valtin, Voss et al., 2007, S. 90). Als Kompromiss werden die Ergebnisse auf Basis von einem *Gesamtscore Lesen* (1D-Modell) und von zwei zweidimensionalen Modellvarianten berichtet. Die 2D-Modelle teilen sich in Leseintentionen (A und B) und Leseverstehensprozesse – *textimmanente Verstehensleistung* wird aus den Verstehensprozessen 1 und 2 gebildet und die *wissensbasierte Verstehensleistung* aus 3 und 4 – auf und besitzen jeweils einen Zusammenhang um 0,9 (Bos, Valtin, Hornberg et al., 2007, S. 150; Bos, Valtin, Voss et al., 2007, S. 92).

| Teilkompetenzen | (1) | (2) | (3) |
|---|---|---|---|
| (1) Elementar phonographisch | | | |
| (2) Elementar grammatisch | 0,89 | | |
| (3) Erweitert phonographisch | 0,87 | 0,88 | |
| (4) Erweitert grammatisch | 0,93 | 0,92 | 0,96 |

Tabelle 4.6: Latente Interkorrelationen des gutschrift-Tests

In IGLU/TIMSS 2011 wurde der Zusammenhang zwischen den zwei Leseintention berechnet, der bei 0,78 liegt (Bos et al., 2012a, S. 238). Höhere Korrelationen weisen die mathematischen und naturwissenschaftlichen Subskalen auf. Sie liegen bei Mathematik innerhalb der drei Skalen *Arithmetik*, *Geometrie/Messen* und *Umgang mit Daten* bei 0,88, 0,83 und 0,89. In der Domäne Naturwissenschaften nehmen die Korrelationen für die Subskalen *Biologie*, *Physik/Chemie* und *Geographie* Werte von jeweils 0,95 an (Bos et al., 2012a, S. 238). Für weitere Analysen wurde in IGLU/TIMSS 2011 aufgrund der ermittelten Korrelationen ein vereinfachtes, restriktiveres dreidimensionales Modell genutzt, welches die Dimensionen Leseverständnis (bzw. -intention), Mathematik und Naturwissenschaften umfasste (Bos et al., 2012a, S. 239). Trotz des hier dargestellten Trends, Subskalen zusammenzufassen, wird bei den Analysen dieser Arbeit die differenzielle Kompetenzstruktur der beiden Tests berücksichtigt, um möglichst vielfältige Informationen über die Eigenschaften der Tests und den orthografischen Kompetenzstand der Kinder zu gewinnen.

Zur Beschreibung der Stärke des Zusammenhangs zwischen den Korrelationen zweier Subskalen soll im Folgenden zudem der *Determinationskoeffizient* als Maßzahl eingeführt werden. Über ihn lassen sich Korrelationen als Effektstärkemaße interpretieren (B. Rasch et al., 2004, S. 121). Er ergibt sich aus dem quadrierten Korrelationskoeffizienten $r_{xy}$, und beschreibt den Anteil der gemeinsamen Varianz beider Merkmale (Rost, 2004, S. 385 f.). Das heißt, je mehr Varianz die beiden Teilkompetenzen gemeinsam haben, desto mehr hängen sie miteinander zusammen und desto weniger ist eine Unterscheidung sinnvoll.

In Tabelle 4.6 sind die latenten Interkorrelationen auf der unteren Dreiecksmatrix dargestellt. Sie liegen im Bereich von 0,87 bis 0,96. Insgesamt betrachtet sind die Werte, verglichen mit den oben beschriebenen Befunden in PISA und IGLU, nicht übermäßig erhöht oder unüblich. Am niedrigsten fallen die Zusammenhänge für die ersten drei Kompetenzen untereinander aus; sie betragen 0,89, 0,87 und 0,88. Der gemeinsame Varianzanteil liegt bei 76 Prozent für die beiden Niveauausprägungen der phonographischen Kompetenz, bei 77 Prozent für die elementare grammatische und die erweiterte phonographische Kompetenz sowie bei 79 Prozent für die beiden voneinander unterschiedenen Kompetenzen auf elementarer Stufe. Der Zusammenhang zwischen den elementaren Teilkompetenzen fällt damit, verglichen mit dem in der Studie zur Frühdiagnose rechtschreibschwächerer Schülerinnen und Schüler getesteten zweidimensionalen Modell, welches eine Korrelation von 0,58 aufwies, deutlich höher aus (vgl. Abschnitt 2.2.2). Allerdings ist dabei zu beach-

ten, dass der Test sich im Rahmen der wissenschaftlichen Begleitforschung auf die zweite Jahrgangsstufe bezieht und daher nur die elementaren Kompetenzen prüft, wohingegen in ZuRecht das gesamte orthografische Kompetenzmodell modelliert und skaliert worden ist.

Um dennoch die Korrelationen der fähigkeitsbasierten Dimensionen zu quantifizieren, wurde neben den Ergebnissen in Tabelle 4.6 eine zusätzliche Skalierung des zweidimensionalen, fähigkeitsbasierten Modells berechnet, welches bereits bei dem Modellvergleich, der zu Beginn dieses Abschnitts beschrieben wurde, Erwähnung fand (vgl. Tabelle 4.4). Dafür wurde das Modell optimiert, sodass 84 Items verblieben. Es wurde ein Zusammenhang von 0,98 zwischen der phonographischen Teilkompetenz und der grammatischen Teilkompetenz quantifiziert. Der Wert übersteigt alle Korrelationen des Modells in Tabelle 4.6. Dieses Ergebnis unterstreicht ebenfalls die Weiterarbeit mit dem vierdimensionalen Modell, das die Basis für alle weiteren Analysen bildet.

Die Stärke des Zusammenhangs erweist sich mit 92 Prozent bei dem 4D-Modell in Tabelle 4.6 zwischen den erweiterten Teilkompetenzen (0,96) als am höchsten. Aus analytischer Sicht ist dies ein Indikator für eine schwache Unterscheidung der beiden Teilfähigkeiten, da sie ähnliche Informationen erfassen. Allerdings fallen die Korrelationen der erweiterten mit den elementaren Kompetenzen unterschiedlich aus und betragen bei der elementaren phonographischen Teilkompetenz 0,87 und 0,93 und bei der elementaren grammatischen Kompetenz 0,88 und 0,92, was einen Grund für eine Unterscheidung darstellen könnte. Die genannte Korrelation von 0,93 zwischen der elementaren phonographischen und der erweiterten grammatischen Teilkompetenz stellt den zweithöchsten ermittelten Zusammenhang dar. Er beträgt 86 Prozent, obwohl es sich um unterschiedliche Fähigkeiten und Niveauausprägungen handelt.

### Arbiträre Itemklassifikation

Um zu zeigen, dass die Ergebnisse der mehrdimensionalen Analyse des gutschrift-Tests nicht zufällig zustande gekommen sind, wurde analog zu Wu und Adams (2006, S. 105 ff.) eine arbiträre Zuordnung der Items zu den Kompetenzdimensionen vorgenommen. Dafür wurden die 82 Items des finalen Raschmodells verwendet und die Items 1, 5, 9, 13, 17 ... 81 der ersten Teilkompetenz zugeordnet, die Items 2, 6, 10, 14, 18 ... 82 der zweiten usw. Die Zuordnung erfolgte damit beliebig, indem sie nicht durch theoretische Vorannahmen gelenkt wurde, die Items aber ausgewogen auf die vier Dimensionen verteilt wurden (die 1. und 2. willkürliche Dimension umfasst jeweils 21 Items und die 3. und 4. Dimension 20 Items). Die Ergebnisse der arbiträren Itemklassifikation finden sich in Tabelle 4.7.

Die Korrelationen fallen durchgängig hoch aus und nähern sich 1 an, was für einen nahezu perfekten Zusammenhang der Dimensionen spricht und deutlich macht, dass die vier Dimensionen nicht unterschieden werden können. Diese korrelativen Strukturen auf bivariater Ebene decken sich mit den Ergebnissen auf multivariater Ebene. Der Deviance-Wert des Modells beträgt 35.188,40 und fällt damit 142 Punkte schlechter aus als für das Modell mit den theoriekonformen Zuordnungen der Items zu den Teilkompetenzen. Er liegt damit fast so hoch wie für das Generalfaktormodell mit einem Wert von 35.192,81

| willkürliche Dimensionen | 1 | 2 | 3 |
|---|---|---|---|
| 1. willkürliche Dimension | | | |
| 2. willkürliche Dimension | 0,98 | | |
| 3. willkürliche Dimension | 0,98 | 0,99 | |
| 4. willkürliche Dimension | 0,97 | 0,98 | 0,98 |

Tabelle 4.7: Latente Interkorrelationen des gutschrift-Tests bei arbiträrer Itemklassifikation

(vgl. Tabelle 4.5). Die Ergebnisse zeigen, dass die Befunde aus Tabelle 4.6 damit nicht allein zufällig zustande kommen können.

Die vorgestellten Analysen der alternativen Modelle weisen – basierend auf den Ergebnissen des Itemreviews, der Tests auf Gesamtmodellebene sowie der korrelationsstatistischen Analyse – das vierdimensionale gutschrift-Kompetenzmodell mit dichotomer Datenstruktur als das Modell mit der höchsten Erklärungskraft für die in den Daten enthaltenden Informationen aus. Daher wird es für alle weiteren Berechnungen genutzt.

### 4.2.2 Sprachsystematischer Rechtschreibtest

#### Itemreview

Beim SRT wurden 261 Items in einem mehrdimensionalen Raschmodell ($n$ = 562) mit fünf Teilkompetenzen bzw. Dimensionen skaliert:

1. Phonographisch-silbisches Prinzip
2. Morphologisches Prinzip
3. Peripheriebereich
4. Prinzip der Wortbildung
5. Wortübergreifendes Prinzip

Insgesamt entsprachen 52 der 261 Items nicht den Kriterien für ein gutes Item und wurden aufgrund eines schlechten Infits und/oder geringer Trennschärfe von dem Test ausgeschlossen.

Im Folgenden ist ein Beispiel für eine Aufgabe mit gutem Fit und mit einem Misfit dargestellt, die anhand der Beobachtung der ICC graphisch veranschaulicht werden kann (DeMars, 2010, S. 10; Furr & Bacharach, 2008, S. 324; Wu & Adams, 2007, S. 65 f.). Mittels der ICC zeigt sich, ob die Items die beobachtete Verteilung auf dem Fähigkeitskontinuum gut abbilden. Der Fit wird durch den Vergleich der empirischen ICC mit der theoretischen ICC geprüft (Embretson & Reise, 2000, S. 234). Die empirische ICC ergibt

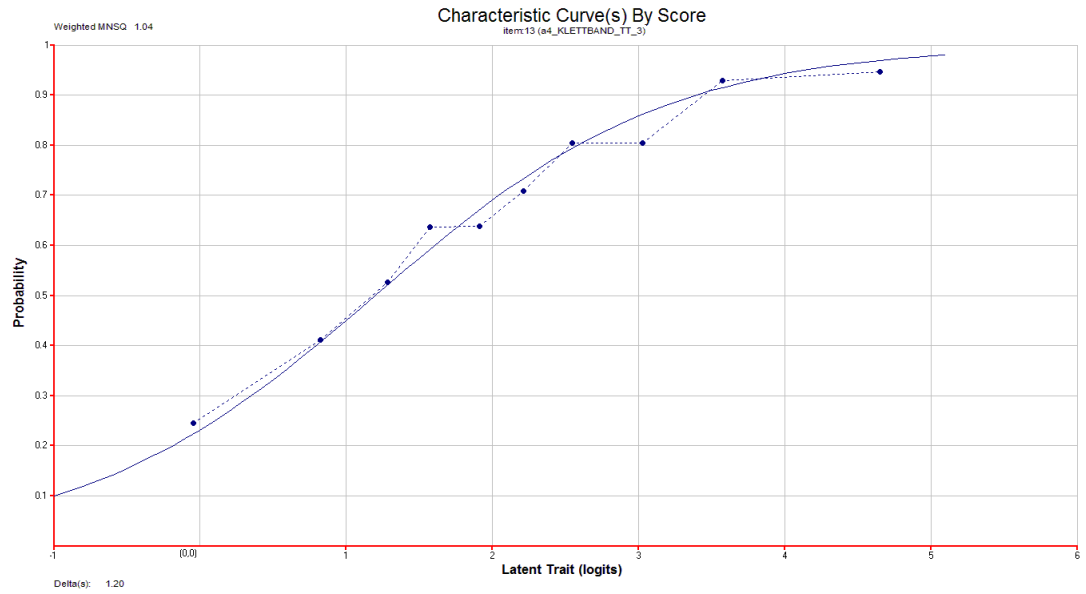



Abbildung 4.3: ICC: Illustration eines Items mit gutem Modellfit

sich aus der a-posteriori-Verteilung der Fähigkeiten für jedes Kind, welches das Item beantwortet hat, und basiert damit auf den direkt beobachteten Daten bzw. dem beobachteten Antwortverhalten (Foy, Galia & Li, 2007, S. 163). Die theoretische Kurve bildet sich aus den Daten der geschätzten Itemparameter; dies sind die vom Raschmodell angenommenen Werte.

Abbildung 4.3 veranschaulicht den beobachteten und den theoretischen Funktionsverlauf des dichotomen Items ⟨tt⟩ des Testwortes Klettband aus einer Skalierung des Peripheriebereichs. Auf der Ordinate kann die Wahrscheinlichkeit zur richtigen Schreibung des Items in Abhängigkeit von der Personenfähigkeit, die auf der Abszisse dargestellt ist, abgelesen werden. In diesem Beispiel hat ein Kind mit einer Personenfähigkeit von 1,2 eine Wahrscheinlichkeit von 50 Prozent, die Struktureinheit ⟨tt⟩ korrekt zu verschriftlichen. Der empirische Funktionsverlauf ist in Abbildung 4.3 als gestrichelte und der theoretisch erwartete Funktionsverlauf als durchgezogene Linie dargestellt. Die beiden ICCs haben einen ähnlichen Verlauf, d. h. für das betrachtete Item ergibt sich eine gute Passung des Modells mit den Daten. Der WMNSQ-Wert von 1,04, also Nahe 1, bestätigt dies.

Im Gegensatz dazu ist in Abbildung 4.4 ein Beispiel für ein Item dargestellt, das während des Itemreviewprozesses bei der Auswertung im Bereich des wortübergreifenden Prinzips ausgeschlossen worden ist. Es handelt sich hierbei um das Wort Malen, bei dem die Großschreibung als Item betrachtet wird. Die Kurven weichen stark voneinander ab: die empirisch beobachtete Kurve ist flacher als die theoretisch vom Raschmodell erwartete. Dies deutet auf einen schwachen Zusammenhang zwischen dem richtigen Schreiben des Items und der wortübergreifenden Teilkompetenz hin. Der flachere Verlauf ist typisch bei einem MNSQ-Wert größer 1 (Wu & Adams, 2007, S. 66; Rost, 2004, S. 374). Der Underfit von 1,43 bestätigt, dass es sich hier um kein gutes Item handelt.

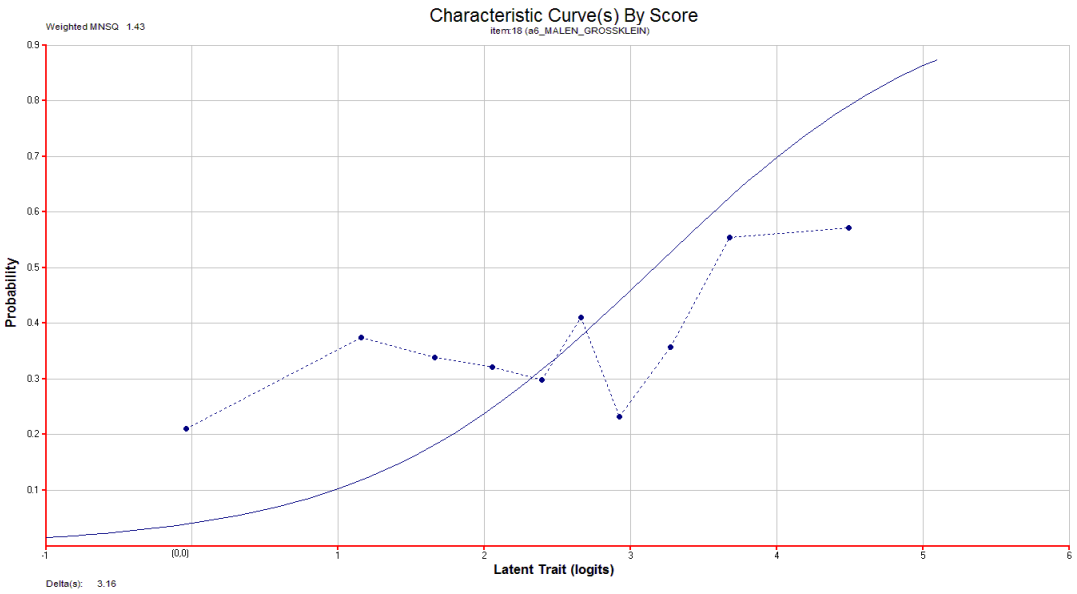



Abbildung 4.4: ICC: Illustration eines Items mit schlechtem Modellfit

Die Wright Map in Abbildung 4.5 zeigt, wie zuvor bei gutschrift, eine linksschiefe Verteilung. Die Parameterverteilungen sind gegeneinander verschoben: Die Personen sind fähiger als die Aufgaben schwer sind. Dementsprechend ist der Test eher zu leicht, aber für eine differenzierte Trennung, insbesondere im unteren Leistungsspektrum, angemessen. Aber auch hier sollten Items mit höherem Schwierigkeitsgrad ergänzt werden. So liegen die mittleren Personenfähigkeiten von 3,102 (Dimension 1), 2,319 (Dimension 2), 2,051 (Dimension 3), 2,783 (Dimension 4) und 2,746 (Dimension 5) deutlich über den mittleren Aufgabenschwierigkeiten von 0,047 (Dimension 1), 0,047 (Dimension 2), 0,060 (Dimension 3), 0,048 (Dimension 4) und 0,067 (Dimension 5) Logits. Die durchschnittliche Streuung fällt mit 1,252 (Dimension 1) und 1,229 (Dimension 2) für die beiden Prinzipien im Kernbereich am kleinsten aus. Für die weiteren Teilkompetenzen nimmt sie Werte von 2,049 (Dimension 3), 1,308 (Dimension 4) und 2,515 (Dimension 5) an. Die Fähigkeitswerte reichen von ca. -1,5 bis 7 und die Itemschwierigkeiten von -3,548 bis 5,121 Logits.

Das schwierigste Item ist Nummer 205 (vgl. Abbildung 4.5), hinter dem sich die Großschreibung des Testwortes Vergnügen, das 16 Prozent der Schülerinnen und Schüler richtig geschrieben haben, verbirgt. Es hat eine Schwierigkeit von 5,121 Logits. Weiterhin erfordert u. a. die richtige Schreibung der Struktureinheiten ⟨rr⟩ in herrliche (Nummer 141 mit einer Schwierigkeit von 3,993 Logits und 19 Prozent Richtigschreibungen) und ⟨schnurr⟩ in Schnurrbarthaaren (Nummer 126, Schwierigkeit 3,782 Logits, 23 Prozent Richtigschreibungen) eine hohe Eigenschaftsausprägung. Das leichteste Item mit einer Schwierigkeit von -3.548 Logits und 99,5 Prozent Richtigschreibungen bildet die Analyseeinheit ⟨a⟩ in dem Testwort Gans (Nummer 95).

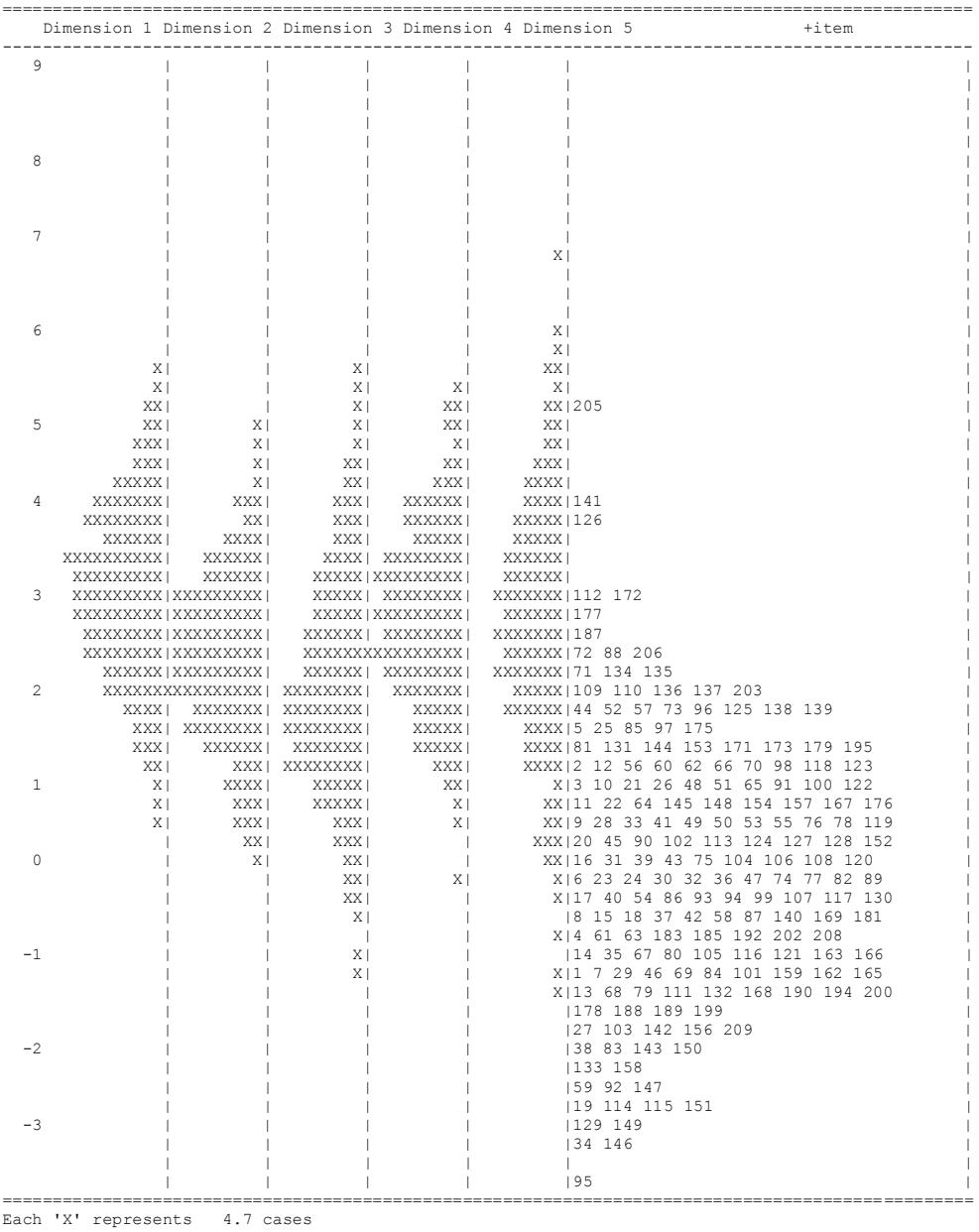



Abbildung 4.5: Wright Map für das fünfdimensionale SRT-Modell

### Gegenüberstellung mit dem Generalfaktormodell

Um die Dimensionalität der Daten empirisch zu prüfen, wurde für den SRT, neben dem mehrdimensionalen Modell mit den fünf der Theorie nach unterschiedenen Teilkompetenzen, ein Generalfaktormodell berechnet, in dem alle Struktureinheiten in einer Kompetenzdimension erfasst werden und das auf denselben Daten fußt. Aus Tabelle 4.8 lässt sich ablesen, dass der Deviance-Wert für das fünfdimensionale Modell im Vergleich zum Generalfaktormodell um 635,81 Punkte niedriger liegt. Auch die Informationsindizes bevorzugen dieses Modell. Sie fallen ebenfalls deutlich niedriger für das differenzierte Kompetenzmodell aus. Die Differenz der Deviance-Werte für das ein- und mehrdimensionale Modell folgt einer $\chi^2$-Verteilung mit 14 Freiheitsgraden. Der kritische $\chi^2$-Wert von 29,14 (df = 14, $\alpha$ = 1%) ist kleiner als der empirische Wert von 635,81. Die Wahrscheinlichkeit für eine Geltung von $L_0$, also des konkurrierenden Generalfaktormodells, ist damit kleiner als 1 Prozent. Das 5D-Modell hat demnach eine bessere Anpassungsgüte und eine höhere Erklärungskraft für die in den Daten enthaltenen Informationen.

### Korrelationsstatistische Analyse

Die latenten Interkorrelationen sind für den SRT in Tabelle 4.9 abgetragen. Die Ergebnisse zeigen eine Übereinstimmung mit den Werten, die bereits in Blatt et al. (2011, S. 246) veröffentlicht sind und wo der Test in unveränderter Form eingesetzt worden ist. Sie sind in drei Fällen identisch und weichen ansonsten nur ab der zweiten Nachkommastelle ab: die Differenz beträgt einmal 0,03, einmal 0,02 sowie fünfmal 0,01 Punkte. Die Kodierung erfolgte dort computerunterstützt mit MaxQda. Die in ZuRecht standardisierte Kodierung mittels SRT-Editor und die weiterführenden neuen Skalierungen weichen davon also nur marginal ab.

Am höchsten korrelieren mit 0,96 die beiden Teilkompetenzen im Kernbereich sowie das phonographisch-silbische Prinzip und das Wortbildungsprinzip miteinander. Der Determinationskoeffizient beträgt 92 Prozent. Bereits bei gutschrift konnte die gleiche Höhe des Zusammenhangs zwischen zwei Subskalen ermittelt werden. Auch hier kann die Eigenständigkeit der Teilkompetenzen nach der korrelationsstatistischen Analyse nicht bestätigt werden. Die Korrelationen des phonographisch-silbischen Prinzips und des morphologischen Prinzips sind darüber hinaus mit dem Peripheriebereich identisch (0,83) und liegen mit den zwei weiteren Teilkompetenzen nahe beieinander (0,96 und 0,95 sowie 0,83 und 0,80). Diese Zusammenhänge der beiden Teilkompetenzen aus dem Kernbereich würden für Hinneys These von der elementaren Wortschreibkompetenz, die sie im zweischrittigen Konstruktionsprinzip (vgl. Abschnitt 2.3.1) umsetzt, sprechen. Das Prinzip der Wortbildung korreliert hingegen unterschiedlich hoch mit dem Peripheriebereich (0,78). Niedrigere latente Korrelationsmuster ergeben sich durchweg mit dem Peripheriebereich und dem wortübergreifenden Prinzip. Aus psychometrischer Sicht können damit die theoretischen Annahmen, die zu einer Ausdifferenzierung in diese beiden Teilkompetenzen geführt haben, bestätigt werden. Die kleinsten korrelativen Zusammenhänge sind zwischen dem

| Modell | Deviance | Parameter | AIC | BIC | CAIC | Δ Deviance | df |
|---|---|---|---|---|---|---|---|
| 5D | 68.749,97 | 224 | 70.392,22 | 70.168,22 | 70.392,22 | 635,81 | 14 |
| 1D | 69.385,78 | 210 | 70.925,39 | 70.715,39 | 70.925,39 | | |

Tabelle 4.8: SRT-Modellvergleich von 5D-Modell und Generalfaktormodell

| Teilkompetenzen | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| (1) Phonographisch-silbisches Prinzip im Kernbereich | | | | |
| (2) Morphologisches Prinzip im Kernbereich | 0,96 | | | |
| (3) Peripheriebereich | 0,83 | 0,83 | | |
| (4) Prinzip der Wortbildung | 0,96 | 0,95 | 0,78 | |
| (5) Wortübergreifendes Prinzip | 0,83 | 0,80 | 0,71 | 0,86 |

Tabelle 4.9: Latente Interkorrelationen des SRT

Peripheriebereich und dem Wortbildungsprinzip von 0,78 sowie dem Peripheriebereich und dem wortübergreifenden Prinzip von 0,71 auszumachen und liegen damit deutlich unter den in PISA ermittelten Koeffizienten in Lesen, Mathematik und Naturwissenschaften (vgl. Abschnitt 4.2.1). Der gemeinsame Varianzanteil beträgt 61 (Dimension 3 und 4) und 50 (Dimension 3 und 5) Prozent. Verglichen mit der Korrelationsmatrix des gutschrift-Tests, in der die geringste Korrelation 0,87 beträgt, können damit unabhängigere Subskalen ausgemacht werden.

Verglichen mit den quantifizierten Zusammenhängen in der IGLU-Voruntersuchung (vgl. Tabelle 2.10) weichen die hier berichteten Ergebnisse folgendermaßen ab: Die Korrelation zwischen dem Peripheriebereich und dem Wortbildungsprinzip ist unverändert, in vier Fällen ist sie um 0,01 bis 0,05 Punkte gesunken und zwischen dem Peripheriebereich und den Kernbereichen sowie zwischen dem Peripheriebereich und dem wortübergreifenden Prinzip um 0,01 bis 0,07 Punkte gestiegen. Die Veränderung des Diktats, infolge der Ergebnisse des Pretests, erzielte für fast alle Werte beim Prinzip der Wortbildung und beim wortübergreifenden Prinzip eine leichte Verbesserung. Eine Verschlechterung zeigt sich ausschließlich bei den Zusammenhängen des Peripheriebereichs mit dem phonographisch-silbischen und dem morphologischen Prinzip. Es kann damit konstatiert werden, dass die korrelative Struktur der Teilkompetenzen des IGLU-Pretests mit den Daten aus ZuRecht reproduziert werden konnte.

| Teilkompetenzen | (1) | (2) | (3) |
|---|---|---|---|
| (1) Phonographisch-silbisches und morphologisches Prinzip im Kernbereich | | | |
| (2) Peripheriebereich | 0,84 | | |
| (3) Prinzip der Wortbildung | 0,95 | 0,78 | |
| (4) Wortübergreifendes Prinzip | 0,82 | 0,70 | 0,84 |

Tabelle 4.10: Latente Interkorrelationen des 4D-Modells des SRT

| Modell | Deviance | Parameter | AIC | BIC | CAIC | Δ Deviance | df |
|---|---|---|---|---|---|---|---|
| 5D | 68.749,97 | 224 | 70.392,22 | 70.168,22 | 70.392,22 | 9,49 | 5 |
| 4D | 68.759,46 | 219 | 69.197,46 | 70.146,06 | 70.365,06 | | |

Tabelle 4.11: SRT-Modellvergleich von 5D-Modell und 4D-Modell

### Modellvergleich: 5D und 4D

Aufgrund des hohen Zusammenhangs zwischen dem phonographisch-silbischen und dem morphologischen Prinzip wurde in ZuRecht zusätzlich ein vierdimensionales Modell berechnet, bei dem diese beiden Teilkompetenzen als eine Dimension skaliert worden sind. Bei der Modelloptimierung verhielten sich die identischen Items wie beim 5D-Modell auffällig und mussten daher eliminiert werden. Die korrelative Struktur ist in Tabelle 4.10 dargestellt und zeigt ähnliche Zusammenhänge. Die Korrelationen der zusammengelegten Kernbereiche mit den weiteren drei Teilkompetenzen weichen um 0,01 Punkte und einmal um 0,02 Punkte im Vergleich zum Modell mit den differenzierten Kernbereichen ab.

Die Statistiken für den Modellvergleich sind in Tabelle 4.11 abgebildet. Der Likelihood-quotiententest zeigt ein nicht signifikantes Ergebnis. Damit erklärt das 4D-Modell die Daten nicht signifikant schlechter als das 5D-Modell und kann als ein mit den Daten verträgliches Modell betrachtet werden. Die Werte sprechen allerdings nicht so stark für das vierdimensionale Modell, wie dies bei dem zuvor berichteten Modellvergleich beim SRT zu beobachten war: Die Prüfgröße von 9,49 liegt nicht so weit entfernt von der Signifikanzgrenze von 15,09, wie dies bei der Gegenüberstellung zwischen 1D- und 5D-Modell (mit einem Wert von 635,81 und einer Grenze von 29,14) zu beobachten war.

Da ein Kompetenzmodell auf fachtheoretisch formulierte Annahmen aufbauen und im Zusammenhang mit der Sprachwissenschaft und -didaktik entwickelt werden muss, soll hier nicht mit dem restriktiveren Modell weitergearbeitet werden. An dieser Stelle kann lediglich darauf aufmerksam gemacht werden, die Modellstruktur weiterzuentwickeln. Die nachfolgenden Analysen beziehen sich daher auf das theoriekonforme 5D-Modell, was eine Betrachtung der Unterschiede in den Werten zwischen dem phonographisch-

| willkürliche Dimensionen | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1. willkürliche Dimension | | | | |
| 2. willkürliche Dimension | 1,00 | | | |
| 3. willkürliche Dimension | 1,00 | 1,00 | | |
| 4. willkürliche Dimension | 1,00 | 1,00 | 0,99 | |
| 5. willkürliche Dimension | 1,00 | 1,00 | 1,00 | 1,00 |

Tabelle 4.12: Latente Interkorrelationen des SRT bei arbiträrer Itemklassifikation

silbischen und dem morphologischen Prinzip ermöglicht, wodurch ebenfalls Hinweise auf die Eigenschaften dieser beiden Teilkompetenzen gewonnen werden können.

#### Arbiträre Itemklassifikation

Das fünfdimensionale Modell für den SRT, in dem die Struktureinheiten beliebig (analog zum zuvor für gutschrift beschriebenen Vorgehen) den Teilkompetenzen zugeordnet wurden, basiert auf den gleichen Daten wie das Modell mit theoriekonformen Annahmen. Es wurde berechnet, um die Befunde in Tabelle 4.9 auf Zufall zu überprüfen. In Tabelle 4.12 sind die Ergebnisse der latenten Korrelationsmuster der arbiträren Itemzuordnung dargestellt. Die Koeffizienten betragen 1 bzw. in einem Fall 0,99, was für einen vollständig linearen Zusammenhang spricht. Damit wird bis zu 100 Prozent der Varianz einer Dimension durch die Varianz der anderen Dimension erklärt. Die formalstatistischen Modellgeltungstests auf der Grundlage der Likelihood-Statistiken weisen dem Modell ebenfalls eine unzureichende Datenanpassung aus: Es hat eine Deviance von 69.384,57 Punkten und damit einen um 634,60 Punkte schlechteren Wert als das fünfdimensionale Modell, in dem die Struktureinheiten theoriebasiert den Teilkompetenzen zugeordnet worden sind. Der Deviance-Wert des Modells mit beliebiger Zuweisung der Items zu den Subskalen fällt damit sehr ähnlich zu dem des Generalfaktormodells aus, der 69.385,78 entspricht (vgl. Tabelle 4.8).

### 4.2.3 Testübergreifende Skalierung

Im Folgenden soll verstärkt auf den Vergleich der beiden Diagnoseinstrumente eingegangen werden. Dafür sei bezüglich des Erhebungsdesigns von ZuRecht nochmals erwähnt, dass es eine Stichprobe von 547 Schülerinnen und Schülern umfasst, für die Testergebnisse aus beiden Rechtschreibtests vorliegen (vgl. Abschnitt 3.1). Dies ermöglicht eine gemeinsame Skalierung der Leistungswerte auf Basis der Kompetenzmodelle von gutschrift und SRT. Das mehrdimensionale Modell weist damit neun unterschiedliche Dimensionen auf. Ein

direkter Vergleich der psychometrischen Struktur der orthografischen Instrumente wird damit ermöglicht. Ergänzend zur theoretischen Analyse der Teilkompetenzen in Abschnitt 2.4 können damit empirisch fundierte Aussagen zu messtheoretischen Gemeinsamkeiten und Unterschieden der Subskalen getroffen werden.

### Itemanzahl und Lösungswahrscheinlichkeit

Für die neundimensionale Skalierung wurden die Items verwendet, die im Rahmen der Itemreviews des vierdimensionalen gutschrift-Modells und des fünfdimensionalen SRT-Modells gute Eigenschaften aufwiesen. Insgesamt beinhaltet die testübergreifende Skalierung damit 291 Aufgaben und setzt sich aus den folgenden Teilkompetenzen bzw. Dimensionen zusammen:

1. Elementar phonographisch
2. Elementar grammatisch
3. Erweitert phonographisch
4. Erweitert grammatisch
5. Phonographisch-silbisches Prinzip
6. Morphologisches Prinzip
7. Peripheriebereich
8. Prinzip der Wortbildung
9. Wortübergreifendes Prinzip

Die mittleren Werte der geschätzten Personenfähigkeiten und Itemschwierigkeiten fallen sehr ähnlich zu den berichteten Werten in den Abschnitten 4.2.1 und 4.2.2 aus. Sie betragen 2,572 und 0,042 (Dimension 1), 2,474 und 0,041 (Dimension 2), 1,937 und 0,061 (Dimension 3), 1,283 und 0,046 (Dimension 4), 3,101 und 0,048 (Dimension 5), 2,291 und 0,047 (Dimension 6), 2,050 und 0,061 (Dimension 7), 2,781 und 0,048 (Dimension 8) sowie 2,744 und 0,067 (Dimension 9) Logits. Die geschätzten durchschnittlichen Lösungswahrscheinlichkeiten der Schülerinnen und Schüler, die durch das Einsetzten der Parameterwerte in Gleichung 3.1 in Abschnitt 3.3.2 errechnet werden können, sind für die Items der erweiterten Teilkompetenzen des gutschrift-Tests am niedrigsten und liegen für die phonographische Dimension bei 87 Prozent und für die grammatische bei 78 Prozent. Die drittkleinste durchschnittliche Lösungswahrscheinlichkeit hat der Peripheriebereich mit 88 Prozent. Die Wahrscheinlichkeit, die Items des phonographisch-silbischen Prinzips richtig zu schreiben, beträgt im Mittel 95 Prozent und ist damit am höchsten. Die weiteren durchschnittlichen Lösungswahrscheinlichkeiten variieren zwischen 90 bis 94 Prozent. Um eine möglichst gute Passung von Itemschwierigkeiten und Personenfähigkeiten zu erhalten, sollten die durchschnittlichen Lösungshäufigkeiten allerdings 50 Prozent betragen (vgl. Abschnitt 3.3.2.2).

| Teilkompetenzen | Anzahl Analyse-einheiten | M | Prozent korrekt | Min. | Max. | SD | VarK |
|---|---|---|---|---|---|---|---|
| Elementar phonographisch | 23 | 20,16 | 88% | 10 | 23 | 2,47 | 12% |
| Elementar grammatisch | 15 | 13,06 | 87% | 5 | 15 | 1,80 | 14% |
| Erweitert phonographisch | 22 | 16,89 | 77% | 1 | 22 | 3,98 | 24% |
| Erweitert grammatisch | 22 | 15,72 | 71% | 3 | 22 | 3,99 | 25% |
| Phonographisch-silbisches Prinzip im Kernbereich | 80 | 72,55 | 91% | 25 | 80 | 7,39 | 10% |
| Morphologisches Prinzip im Kernbereich | 52 | 42,58 | 82% | 8 | 52 | 6,64 | 16% |
| Peripheriebereich | 23 | 17,04 | 74% | 7 | 23 | 3,91 | 23% |
| Prinzip der Wortbildung | 25 | 21,71 | 87% | 4 | 25 | 2,81 | 13% |
| Wortübergreifendes Prinzip | 29 | 24,33 | 84% | 5 | 29 | 4,28 | 18% |

Tabelle 4.13: Univariate Beschreibungen der gutschrift- und SRT-Subskalen

### Univariate Statistiken für die Subskalen

Tabelle 4.13 gibt einen Überblick über die Lage- und Streuungsmaße der vier Subskalen des gutschrift-Tests und der fünf Subskalen des SRT. Laut Adams et al. (1997, S. 11) sind für präzise Schätzungen 20 Items pro Dimension ausreichend. Insgesamt umfasst der gutschrift-Test 82 Struktureinheiten, die sich relativ ausgewogen auf die Teilkompetenzen aufteilen. Die 22 bzw. 23 Items ermöglichen daher konsistente und präzise Schätzungen der Personenfähigkeiten, wobei die elementare grammatische Subskala nur 15 Items umfasst und bei zukünftigen Testkonstruktionen daher mit neuen Items angereichert werden sollte. Der SRT beinhaltet 209 Struktureinheiten. Alle Teilkompetenzen umfassen im Vergleich zum gutschrift-Test eine gleich hohe oder höhere Anzahl an Items. Insbesondere wird die Erhebung der Kernbereiche über viele Items gemessen. Der Peripheriebereich wird im Gegensatz dazu „nur“ über 23 Items operationalisiert. Da es sich der Theorie nach um Ausnahmen und nichtnative Schreibungen handelt, die im Wortschatz seltener vertreten sind, ist der Unterschied in der Anzahl plausibel.

Zunächst zeigt sich, dass die durchschnittlichen Leistungen in den Teilkompetenzen zumeist über den Leistungen auf Ganzwortebene liegen sowie homogener ausfallen (vgl. Tabelle 4.1). Durch Tabelle 4.13 lassen sich Fehlerschwerpunkte der Schülerinnen und Schüller differenzierter erkennen. Die Schreibung der Indikatoren der erweiterten gutschrift-Kompetenzen bereitet den Kindern theoriekonform mehr Probleme als die Schreibung der elementaren Kompetenzen. Die erweiterte grammatische Teilkompetenz ist testübergreifend mit einem Anteil an Richtigschreibungen von 71 Prozent sogar die schwierigste Subskala. Darauf folgt der Peripheriebereich mit einem Prozent-korrekt-Wert von 74 und die erweiterte phonographische Kompetenz mit 77 Prozent. Am leichtesten fällt den Schülerinnen und Schülern aus ZuRecht das phonographisch-silbische Prinzip, in dem sie im Mittel 91 Prozent der Struktureinheiten richtig schreiben. Der Anteil von Falsch-

schreibungen ist in der elementaren phonographischen Teilkompetenz am zweitniedrigsten (12 Prozent). Die Prozent-korrekt-Werte der Teilkompetenzen liegen damit insgesamt zwischen 71 und 91 Prozent. Dies verdeutlicht, dass die Schreibungen der Struktureinheiten von der Mehrheit der Schülerinnen und Schüler gelöst werden.

Die Leistungsvariation ist in den erweiterten Teilkompetenzen und im Peripheriebereich am höchsten. Sie beträgt für die erweiterten Kompetenzen 25 Prozent (grammatisch) und 24 Prozent (phonographisch), sowie für den Peripheriebereich 23 Prozent. Dies ist auch anhand der breiten Verläufe der entsprechenden Personenverteilungen in den Wright Maps (vgl. Abbildungen 4.2 und 4.5) erkennbar. Die Streuung fällt im phonographisch-silbischen Prinzip am geringsten aus (VarK = 10 Prozent). Hier zeigt sich, dass die Kompetenzwerte innerhalb dieses Prinzips homogen sind und es von der Mehrheit der Schülerinnen und Schüler beherrscht wird, während bei den erweiterten gutschrift-Kompetenzen und im Peripheriebereich eine größere Diskrepanz zwischen leistungsschwachen und -starken Rechtschreibern besteht. In allen Subskalen gibt es mindestens ein Kind, das alle Indikatoren der jeweiligen Skala richtig geschrieben hat. Mit nur einer richtig geschriebenen Struktureinheit schneidet ein Kind in der erweiterten phonographischen Kompetenz am schlechtesten ab.

Die Angaben zu gutschrift können zwar nicht unmittelbar mit den Befunden aus IGLU-E 2001 verglichen werden, aber ein indirekter Vergleich wird dennoch durchgeführt, da es sich auch um ein gutschrift-Testinstrument handelt, das ebenso für die vierte Jahrgangsstufe konzipiert worden ist und einen Anteil von einem Drittel identischer Testwörter umfasst. In IGLU 2001 sind ebenfalls die prozentualen Anteile an Richtigschreibungen dokumentiert (vgl. Abschnitt 2.2.2). Die Prozent-korrekt-Werte betragen hier 88 für die elementare phonographische, 93 für die elementare grammatische, 67 für die erweiterte phonographische und 71 Prozent für die erweiterte grammatische Kompetenz. Ebenfalls zeigt sich ein höherer Schwierigkeitsgrad bei den erweiterten Teilfähigkeiten. Indikatoren mit den größten Fehleranteilen in den beiden erweiterten Teilfähigkeiten sind in IGLU-E Dehnungs- und Kürzezeichen, Großschreibung abstrakter Nomen und Nominalisierung, erweiterte Ableitungsoperationen sowie Deklination. Daher ist im Rahmen von ZuRecht wahrscheinlich ebenfalls von vermehrten Fehlern in diesen Bereichen auszugehen, die die Höhe der Prozent-korrekt-Werte in der erweiterten phonographischen und der erweiterten grammatischen Teilkompetenz erklären könnten. Das Abschneiden der Schülerinnen und Schüler in der elementaren phonographischen und der erweiterten grammatischen Subskala ist in IGLU-E und ZuRecht identisch. Die größte Schwierigkeit zeigt sich in IGLU, abweichend von ZuRecht, nicht bei der erweiterten grammatischen, sondern bei der erweiterten phonographischen Kompetenz, die 10 Prozentpunkte niedriger ausfällt. Die zweitgrößte Differenz betrifft die elementare grammatische Teilkompetenz mit 6 Prozentpunkten Unterschied, bei der die Schülerinnen und Schüler aus IGLU 2001 besser abschneiden als die Kinder aus ZuRecht.

Die Resultate des SRT können mit denen aus der IGLU-Voruntersuchung (vgl. Abschnitt 2.3.2) verglichen werden. Auch wenn der SRT in ZuRecht eine Weiterentwicklung des SRT im Pretest ist, so beziehen sich doch beide auf den vierten Grundschuljahrgang

| Teilkompetenzen | (1) Elementar phonographisch | (2) Elementar grammatisch | (3) Erweitert phonographisch | (4) Erweitert grammatisch |
|---|---|---|---|---|
| (5) Phonographisch-silbisches Prinzip im Kernbereich | 0,89 | 0,90 | 0,85 | 0,89 |
| (6) Morphologisches Prinzip im Kernbereich | 0,90 | 0,91 | 0,93 | 0,94 |
| (7) Peripheriebereich | 0,70 | 0,79 | 0,77 | 0,75 |
| (8) Prinzip der Wortbildung | 0,87 | 0,91 | 0,86 | 0,92 |
| (9) Wortübergreifendes Prinzip | 0,79 | 0,83 | 0,71 | 0,83 |

Tabelle 4.14: Latente Interkorrelationen des testübergreifenden Modells

und weisen 76 identische Testwörter sowie neun Wörter, die in einer anderen flektierten oder abgeleiteten Wortform vorkommen, auf. Die Schwierigkeitsreihenfolge ist ähnlich gestuft und geht vom phonographisch-silbischen Prinzip (90 Prozent Richtigschreibungen) über das Prinzip der Wortbildung und das wortübergreifende Prinzip (jeweils 84 Prozent), gefolgt vom morphologischen Prinzip (82 Prozent), schließlich zu dem Peripheriebereich (78 Prozent). Auch die Spannbreite der Leistung ist im Peripheriebereich mit 23 Prozent am größten und im phonographisch-silbischen Prinzip mit 11 Prozent am niedrigsten. Die Prozent-korrekt-Werte und der Variationskoeffizient unterscheiden sich kaum von den Zahlen in ZuRecht; die maximale Differenz beträgt 4 Prozentpunkte.

### Korrelations- und Reliabilitätsanalyse

Bei der korrelationsstatistischen Analyse wird die vier-mal-fünf Matrix fokussiert, die in Tabelle 4.14 dargestellt ist, und in der die Zusammenhänge der differenziellen Teilkompetenzen des gutschrift-Tests (Spalten) und des SRT (Zeilen) abgetragen sind. Wie in Abschnitt 4.2.2 gezeigt worden ist, korrelieren die beiden Prinzipien im Kernbereich des SRT hoch miteinander und weisen identische bzw. sehr ähnliche Zusammenhänge zu den weiteren drei Teilkompetenzen auf. Tabelle 4.14 kann entnommen werden, dass sie mit den erweiterten gutschrift-Subskalen unterschiedlich hoch korrelieren, wobei der Zusammenhang des morphologischen Prinzips stärker ausfällt (Dimensionen 3 und 6: 0,93; Dimensionen 4 und 6: 0,94) als der des phonographisch-silbischen (Dimensionen 3 und 5: 0,85; Dimensionen 4 und 5: 0,89). Der Zusammenhang des morphologischen Prinzips zur erweiterten grammatischen Kompetenz erweist sich innerhalb des gesamten neundimensionalen Modells mit einem gemeinsamen Varianzanteil von 88 Prozent als am höchsten. Dies ist nicht überraschend, da beide Teilkompetenzen ähnliche Indikatoren umfassen, indem sie hauptsächlich die Beherrschung von Schreibungen in flektierten Wörtern testen (vgl. Abschnitt 2.4). Der ebenfalls hohe Zusammenhang des morphologischen Prinzips zur erweiterten phonographischen Kompetenz ist aus theoretischer Sicht ebenfalls plausibel. In der erweiterten phonographischen Teilkompetenz werden die Setzung von doppelten Konsonantengraphemen und des Dehnungs-e in nicht flektierten Wörtern, von doppelten

Vokalgraphemen sowie von Dehnungs-h betrachtet. Die Schreibung dieser Analyseeinheiten stellen zwar keine Struktureinheiten im morphologischen Prinzip dar, die Setzung von Schärfungs- und Dehnungszeichen erfordert aber je nach Ansatz die Fähigkeit der Analyse von Morphemen.

Die geringste Korrelation von 0,70 erzeugen die elementare phonographische Kompetenz und der Peripheriebereich. Der niedrige gemeinsame Varianzanteil von 49 Prozent ist inhaltlich plausibel, da es laut Testkonzeption in der gutschrift-Teilkompetenz insbesondere um Phonem-Graphem-Korrespondenzen geht, während im Peripheriebereich Ausnahme- und Fremdwortschreibungen betrachtet werden. Ebenfalls sind alle weiteren Korrelationen des Peripheriebereichs mit den gutschrift-Teilkompetenzen gering und nehmen Werte von 0,79, 0,77 und 0,75 an. Sie fallen damit durchgängig niedriger als die Zusammenhänge innerhalb des theoriekonformen gutschrift-Kompetenzmodells aus. Zudem unterschreiten sie teilweise die Korrelationen innerhalb der SRT-Kompetenzmodellierung, die bei 0,83, 0,83, 0,78 und 0,71 liegen. Vergleichsweise niedrige Zusammenhangsstrukturen lassen sich darüber hinaus zwischen dem wortübergreifenden Prinzip und den gutschrift-Subskalen beobachten. Der theoretischen Ausdifferenzierung des Peripheriebereichs und des wortübergreifenden Prinzips sowie der Integration in ein Testmodell ist damit Beachtung zu schenken. Die korrelationsspezifische Analyse stellt hier insbesondere eine Eigenständigkeit des Peripheriebereichs heraus, die sich nicht nur im SRT selbst bewiesen hat, sondern auch innerhalb der Skalierung mit den Subskalen von gutschrift geltend ist.

Die phonographischen Kompetenzdimensionen von gutschrift korrelieren mit dem Wortbildungsprinzip des SRT niedriger (Dimensionen 1 und 8: 0,87; Dimensionen 3 und 8: 0,86) im Vergleich zu den grammatischen Dimensionen (Dimensionen 2 und 8: 0,91; Dimensionen 4 und 8: 0,92). Der Determinationskoeffizient zwischen der elementaren bzw. der erweiterten phonographischen Kompetenz und dem Wortbildungsprinzip liegt bei 76 bzw. 74 Prozent. Hier gibt es mit Blick auf Zuordnung von Rechtschreibphänomenen zu den Teilkompetenzen keine Überschneidungen. Der statistische Befund unterstreicht damit die theoretische Gegenüberstellung (vgl. Abschnitt 2.4). Die Korrelationen der grammatischen Subskalen mit dem Prinzip der Wortbildung von über 0,9 sind ebenfalls inhaltlich nachvollziehbar. Alle drei Teilkompetenzen betrachten Wortbildungsprozesse und verfügen daher über Indikatoren, die Wortbildungselemente (Erkennen und Schreiben von Affixen) und das Zusammenführen von Morphemen thematisieren.

Eine wichtige Teststatistik, die für die beiden Orthografietests nachstehend angegeben wird, ist die Reliabilität: „The reliability of an instrument is often used to judge the overall quality of the instrument.“ (Wu & Adams, 2007, S. 69) Sie drückt aus, mit welcher Genauigkeit die Rechtschreibleistungen der Schülerinnen und Schüler über die Teilkompetenzen gemessen werden, d. h. mit welcher Zuverlässigkeit der beobachtete Wert den wahren Wert abbildet. Die Reliabilitäten werden über Korrelationen von Plausible-Value-Ziehungen ermittelt (EAP/PV = Expected A Posteriori/Plausible Value) (Adams & Carstensen, 2002, S. 152).[8] Der Reliabilitätskoeffizient ist ähnlich zu Cronbachs Alpha zu interpretieren (Rost, 2004, S. 382). Er kann Werte zwischen 0 und 1 annehmen (Bühner, 2011, S. 51). Je

[8] Weitere Informationen zu dem Verfahren finden sich z. B. in Adams (2005) und Rost (2004, S. 382).

| Teilkompetenzen | Reliabilität |
|---|---|
| Elementar phonographisch | 0,85 |
| Elementar grammatisch | 0,87 |
| Erweitert phonographisch | 0,90 |
| Erweitert grammatisch | 0,92 |
| Phonographisch-silbisches Prinzip im Kernbereich | 0,90 |
| Morphologisches Prinzip im Kernbereich | 0,92 |
| Peripheriebereich | 0,87 |
| Prinzip der Wortbildung | 0,90 |
| Wortübergreifendes Prinzip | 0,86 |

Tabelle 4.15: Reliabilitäten des testübergreifenden Modells

höher der Wert ausfällt, desto weniger sind die Testunterschiede zwischen Probanden von Messfehlern verzerrt. Die Varianz der Messwerte kann also umso mehr auf tatsächliche Personenunterscheide zurückgeführt werden. Ein Wert von 1 zeigt eine perfekt genaue Messung an. Geringe Werte reduzieren die Aussagekraft der Messung, da nur wenig über die „wahre“ Fähigkeitsausprägung in Erfahrung gebracht wird. Als akzeptabel werden in der Literatur unterschiedliche Werte bezeichnet, mehrheitlich werden aber Koeffizienten ab 0,70 oder 0,80 angegeben (Bortz & Döring, 2006, S. 199; Schermelleh-Engel, Kelava & Moosbrugger, 2006, S. 421).

Die Reliabilitäten (EAP/PV) der Teilkompetenzen der beiden Tests sind in Tabelle 4.15 zusammengetragen. Aufgrund des gemeinsamen Skalierungslaufs können sie direkt aufeinander bezogen werden. Alle Reliabilitäten liegen über 0,8 und sind somit als zufriedenstellend zu bewerten. Die zuverlässigsten Messungen der Personenfähigkeiten erlauben die erweiterte grammatische Teilkompetenz und das morphologische Prinzip mit Werten von jeweils 0,92. Daraufhin weisen mit jeweils 0,90 die erweiterte phonographische Kompetenz, das phonographisch-silbische sowie das Wortbildungsprinzip äußerst gute Reliabilitätsmaßzahlen auf. Insgesamt betrachtet sind die Teilkompetenzen von gutschrift und dem SRT ähnlich reliabel und damit für die Erfassung der orthografischen Kompetenz angemessen. Es kann auf Basis der Werte in Tabelle 4.15 kein Test identifiziert werden, der genauere Messungen erlaubt als der andere.

### 4.2.4 Einbindung der IGLU-Hauptuntersuchung

Der SRT wurde unverändert – mit den gleichen Testwörtern und Struktureinheiten – ebenfalls in der IGLU-Hauptuntersuchung 2006 eingesetzt. Da hierzu bisher keine veröffentlichten Befunde existieren, wurden die Daten im Rahmen dieser Arbeit analysiert. Insgesamt liegen zu 2.549 Schülerinnen und Schülern aus ganz Deutschland auswertbare Ergebnisse zu dem Diktat vor. Um das Kompetenzniveau der Kinder aus ZuRecht einordnen zu können, soll dieses in Beziehung zu der repräsentativen Haupterhebung (HE) gesetzt werden. Damit wird untersucht, ob die ZuRecht-Stichprobe eher unter-, über- oder durchschnittliche Rechtschreibkompetenzen aufweist. Ferner kann geprüft werden, ob die Befunde zur Kompetenzmodellierung aus ZuRecht durch die Ergebnisse der HE gestützt werden. Daher sollen im Folgenden größtenteils die gleichen Statistiken, wie sie zuvor für ZuRecht in Abschnitt 4.2.2 dargestellt worden sind, berichtet werden, um sie gegeneinander zu halten. Eine Vergleichbarkeit ist dadurch gewährleistet, dass die Dateneingabe und die Kodierung der Testdaten in der HE nach identischen Regularitäten und Vorgehensweisen, wie sie in Abschnitt 3.2 für den SRT beschrieben wurden, erfolgten. Die Daten der HE wurden mittels *Student House Weight* gewichtet, sodass die Analysen für die Schülerinnen und Schüler der Population repräsentativ sind.[9]

#### Itemreview

Bei dem mehrdimensionalen Modell des SRT in der HE stellen die fünf Teilkompetenzen erneut die folgenden Dimensionen dar:

1. Phonographisch-silbisches Prinzip
2. Morphologisches Prinzip
3. Peripheriebereich
4. Prinzip der Wortbildung
5. Wortübergreifendes Prinzip

Von den 261 Items des fünfdimensionalen Skalierungslaufes haben sich 220 Items als modellkonform erwiesen. Damit wurden 11 Items weniger aus dem Datensatz aufgrund schlechter Eigenschaften entfernt als in ZuRecht (25 Items, die in der Haupterhebung eliminiert worden sind, sind auch in ZuRecht entfernt worden). Die Verteilung der Itemschwierigkeiten in Bezug zu den Personenfähigkeiten in Form einer Wright Map ist in Abbildung 4.6 visualisiert. Auch hier zeigt sich, dass sich die Items verstärkt im mittleren bis unteren Bereich der Skala ansiedeln und daher eher zu leicht für die Schülerinnen und Schüler sind. Viele Items werden mit einer sehr hohen Wahrscheinlichkeit gelöst, was sich anhand der Mittelwerte der Personen- und Aufgabenparameter zeigt: Sie betragen 3,421 und 0,033 für die erste Dimension, 2,868 und 0,036 für die zweite Dimension, 2,401 und

[9] gutschrift wurde ebenfalls unverändert in der HE eingesetzt, konnte hier aber nicht ausgewertet werden, da die dafür benötigten Kodierungen bisher nicht von den Testautorinnen zur Verfügung gestellt wurden.

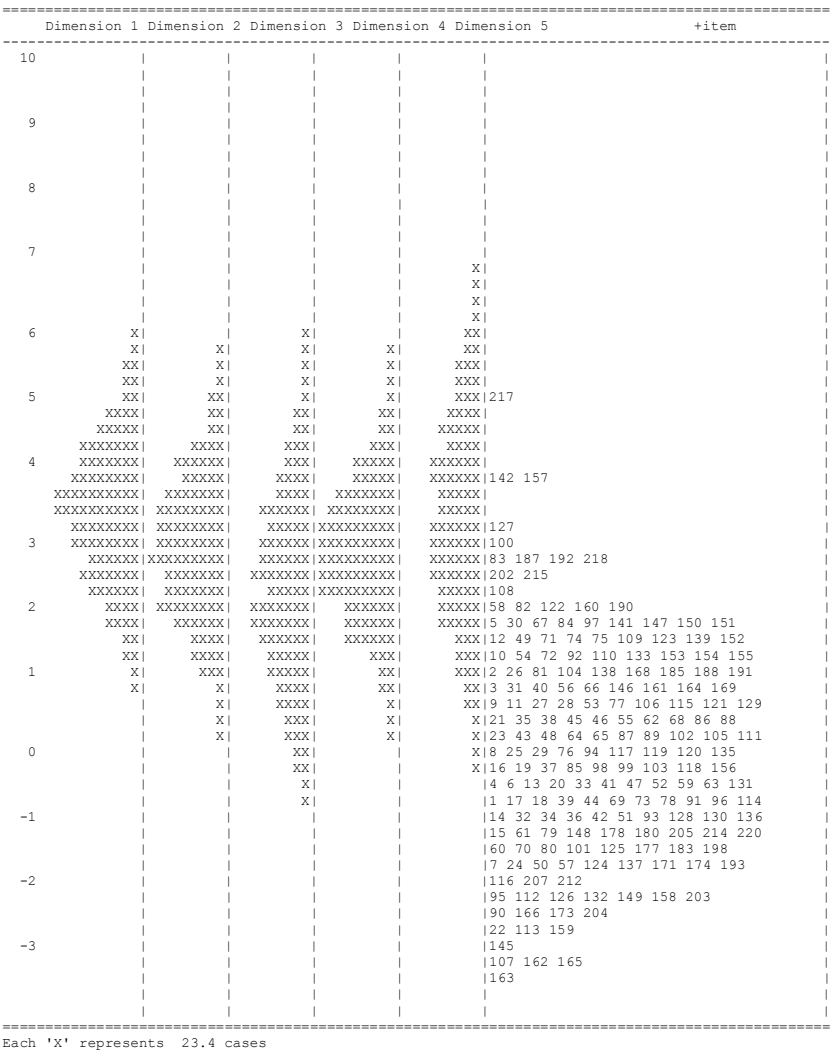



Abbildung 4.6: Wright Map für das fünfdimensionale SRT-Modell in der IGLU-HE

0,046 für die dritte Dimension, 2,848 und 0,034 für die vierte Dimension sowie 3,255 und 0,051 für die fünfte Dimension. Die Streuung nimmt Werte von 1,266 (Dimension 1), 1,476 (Dimension 2), 2,462 (Dimension 3), 1,327 (Dimension 4) und 2,922 (Dimension 5) an.

### Gegenüberstellung mit dem Generalfaktormodell

Die Resultate des Modellvergleichs zwischen dem mehrdimensionalen Modell, das theoriekonform ist, und dem eindimensionalen Modell, das Rechtschreibung als globale Fähigkeit definiert, sind in Tabelle 4.16 wiedergegeben. Die Befunde aus ZuRecht können an dieser Stelle ebenfalls reproduziert werden. Auch hier weisen die Informationsindizes und der $\chi^2$-Test das fünfdimensionale als das Modell aus, welches die Daten besser erklärt. Der

| Modell | Deviance | Parameter | AIC | BIC | CAIC | Δ Deviance | df |
|---|---|---|---|---|---|---|---|
| 5D | 130.555,42 | 235 | 130.995,42 | 132.280,98 | 132.500,98 | 1.295,35 | 14 |
| 1D | 131.850,78 | 221 | 132.290,78 | 133.576,34 | 133.796,34 | | |

Tabelle 4.16: SRT-Modellvergleich von 5D-Modell und Generalfaktormodell für die IGLU-HE

| Teilkompetenzen | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| (1) Phonographisch-silbisches Prinzip im Kernbereich | | | | |
| (2) Morphologisches Prinzip im Kernbereich | 0,96 | | | |
| (3) Peripheriebereich | 0,82 | 0,85 | | |
| (4) Prinzip der Wortbildung | 0,90 | 0,92 | 0,82 | |
| (5) Wortübergreifendes Prinzip | 0,80 | 0,82 | 0,73 | 0,87 |

Tabelle 4.17: Latente Interkorrelationen des SRT in der IGLU-HE

errechnete Wert der Prüfgröße (1.295,35) liegt weit oberhalb der kritischen Grenze von 29,14 (df = 14, $\alpha$ = 1%). Demnach ist das 5D-Modell gegenüber dem Generalfaktormodell vorzuziehen.

### Korrelationsstatistische Analyse

Die latenten Korrelationen in der HE (vgl. Tabelle 4.17) sind vergleichbar mit den Zusammenhängen in Tabelle 4.9 aus ZuRecht. Sie unterscheiden sich um maximal 0,06 Punkte. Diese Differenz betrifft die Korrelation zwischen dem phonographisch-silbischen Prinzip und dem Prinzip der Wortbildung, die hier von 0,96 auf 0,90 gesunken ist. Wie bereits bei ZuRecht beobachtet, fallen zudem die Zusammenhänge des Wortbildungsprinzips zu den Subskalen der beiden Kernbereiche, des Peripheriebereichs und des wortübergreifendes Prinzips unterschiedlich hoch aus: Sie betragen für die beiden Kernbereiche 0,90 und 0,92, für den Peripheriebereich 0,82 und für das wortübergreifende Prinzip 0,87. Die Beobachtungen zum Prinzip der Wortbildung aus ZuRecht und der IGLU-HE können als Argument für die eigenständige Betrachtung des Kompetenzbereiches herangezogen werden. Die Korrelation zwischen den Kernbereichen ist mit 0,96 hingegen identisch hoch. Damit stellen auch die Analysen im Rahmen der IGLU-HE die Unterscheidung des phonographisch-silbischen und des morphologischen Prinzips in Frage. Die Korrelationen des Peripheriebereichs und des wortübergreifenden Prinzips mit den weiteren Subskalen liegen in der HE leicht über bzw. unter den ermittelten Zusammenhängen in ZuRecht. Der Varianzanteil liegt zwischen 53 und 76 Prozent. In der HE sind damit ebenfalls diese beiden Teilkompetenzen unter analytischen Gesichtspunkten am eigenständigsten.

| Wortebene/ Teilkompetenzen | Anzahl Analyse-einheiten | M | Prozent korrekt | Min. | Max. | SD | VarK |
|---|---|---|---|---|---|---|---|
| Wortebene | 121 | 94,77 | 78% | 6 | 120 | 15,51 | 16% |
| Wortebene selektiert | 78 | 53,27 | 68% | 1 | 77 | 14,16 | 27% |
| Phonographisch-silbisches Prinzip im | 91 | 83,99 | 92% | 3 | 91 | 7,46 | 9% |
| Morphologisches Prinzip im Kernbereich | 57 | 48,84 | 86% | 3 | 57 | 6,62 | 14% |
| Peripheriebereich | 22 | 16,81 | 76% | 0 | 22 | 3,93 | 23% |
| Prinzip der Wortbildung | 25 | 21,63 | 87% | 2 | 25 | 2,99 | 14% |
| Wortübergreifendes Prinzip | 25 | 21,39 | 86% | 1 | 25 | 3,61 | 17% |

Tabelle 4.18: Univariate Beschreibungen der SRT-Subskalen in der IGLU-HE

Univariate Statistiken für die Subskalen

Sowohl bei der Auszählung der Analyseeinheiten auf Gesamtwortebene als auch für die Teilkompetenzen lässt sich erkennen, dass die Schülerinnen und Schüler aus der HE (vgl. Tabelle 4.18) vergleichbare Werte wie die Kinder aus ZuRecht (vgl. Tabellen 4.1 und 4.13) erzielen. Insgesamt schneiden die IGLU-Kinder leicht besser ab: Die Prozent-korrekt-Werte fallen in der HE durchschnittlich um rund 2 Prozentpunkte höher aus. Die maximale Differenz zeigt sich auf Wortebene und im morphologischen Prinzip mit jeweils 4 Punkten, wobei auch dieser Unterschied verhältnismäßig klein ist. Bei dem Wortbildungsprinzip findet sich keine Differenz. Die Werte des Variationskoeffizienten sind ebenfalls sehr ähnlich zu den zuvor berichteten (vgl. Tabelle 4.13). Sie unterscheiden sich um höchstens 2 Prozentpunkte. Die maximale Anzahl an richtig geschriebenen Struktureinheiten wurde, analog zum Einsatz in ZuRecht, von jeweils mindestens einem Kind realisiert. Das schlechteste Ergebnis für eine Teilkompetenz weist ein Viertklässler mit keiner richtig geschriebenen Struktureinheit im Peripheriebereich auf. Auch in der Hauptuntersuchung wurde das ganze Diktat des SRT nicht ein Mal vollständig richtig verschriftet. Das beste Ergebnis erzielen fünf Schülerinnen und Schüler mit 120, das schlechteste Ergebnis ein Kind mit 6 normgerecht geschriebenen Wörtern.

### 4.2.5 Zusammenschau

In diesem Abschnitt wurden die Eigenschaften der Rechtschreibtests gutschrift-diagnose und SRT sowie die orthografische Leistung der Schülerinnen und Schüler über differenzierte Angaben zu den Teilkompetenzen dargestellt und die empirische Validität geprüft. Es wurde dabei zunächst gezeigt, dass die Daten des gutschrift-Tests ohne größeren Informationsverlust mit dem dichotomen Raschmodell ausgewertet werden können. Da das „einfachere“ Modell (Einfachheitskriterium) die Daten sogar besser erklärt, basieren alle weiteren Analysen auf diesem Raschmodell. Die für beide Tests anschließenden multiva-

riaten Dimensionalitätsanalysen zeigten ein signifikantes Ergebnis: Es konnte eine bessere Anpassung der theoriekonformen mehrdimensionalen Modelle an die Datenstruktur im Vergleich zu den Generalfaktormodellen nachgewiesen werden. Demzufolge ergibt sich eine empirische Evidenz für die differenziellen Rechtschreibkompetenzmodelle.

Die korrelationsstatistischen Analysen bestätigten teilweise diesen Befund. Einige Subskalen weisen sehr hohe Zusammenhänge auf. Insbesondere sind hier für gutschrift die erweiterten Teilkompetenzen und für den SRT die beiden Prinzipien im Kernbereich zu nennen. Hier sollte auf Basis der psychometrischen Kennwerte zukünftig die Struktur überdacht werden. Die theoretische Ausdifferenzierung des Peripheriebereichs und des wortübergreifenden Prinzips des SRT sind hingegen aus analytischer Sicht bestätigt worden. Sie weisen, verglichen mit den gutschrift-Subskalen, durchgängig niedrigere Korrelationen auf. Die Auswertung der Daten aus der IGLU-Hauptuntersuchung bestärkt die in ZuRecht vorgefundenen Zusammenhänge zwischen den SRT-Subskalen, da sich durchweg vergleichbare Korrelationen beobachten ließen.

Die testübergreifende Skalierung konnte auf bivariater Ebene ebenfalls die Selbstständigkeit des Peripheriebereichs und des wortübergreifenden Prinzips hervorheben, indem diese die niedrigsten Zusammenhänge zu den gutschrift-Teilkompetenzen aufwiesen. Zudem wurde die theoretische Gegenüberstellung der beiden Tests aus Abschnitt 2.4, nach der z. B. die erweiterte grammatische Kompetenz und das morphologische Prinzip eine große Menge ähnlicher Indikatoren aufweisen, oder die elementare phonographische Kompetenz und der Peripheriebereich kaum gleiche Struktureinheiten umfassen, durch die Korrelationsmatrix größtenteils statistisch verifiziert. Die Reliabilitäten, die auf Basis des gemeinsamen Modelllaufs unmittelbar in Beziehung zueinander gesetzt werden konnten, ergaben vergleichbar gute Reliabilitäten für gutschrift und den SRT.

Die neundimensionale Skalierung verdeutlichte auf univariater Ebene, dass beide Erhebungsinstrumente einen ähnlichen Schwierigkeitsgrad aufweisen und eher zu leicht sind, aber auch jeweils Teilkompetenzen aufweisen – damit sind die zwei erweiterten gutschrift-Teilkompetenzen und der Peripheriebereich gemeint – deren Lösungswahrscheinlichkeiten und Prozent-korrekt-Werte theoriekonform deutlich unter denen der anderen Subskalen liegen. Für beide Tests wäre es demzufolge sinnvoll, sie um weitere, schwierigere Items anzureichern, um auch die Schülerinnen und Schüler im oberen Leistungsbereich differenziert erfassen zu können. Bei dem gutschrift-Test sollten zudem weitere Aufgaben für die elementare grammatische Subskala ergänzt werden, da sich insgesamt nur 15 Items auf Basis der Itemselektionskriterien als modellkonform erwiesen haben. Die parallelen deskriptiven Auswertungen in IGLU für den SRT ergaben minimale Abweichungen der Häufigkeitsverteilungen auf Gesamtwort- und Teilkompetenzebene. Der Kompetenzstand der Kinder aus ZuRecht unterscheidet sich demnach nur marginal von dem der IGLU-HE, womit gezeigt werden konnte, dass die in ZuRecht getesteten Schülerinnen und Schüler ein durchschnittliches Kompetenzniveau aufweisen.

## 4.3 Zusammenhänge mit Hintergrundmerkmalen

In diesem Abschnitt werden die erhobenen Hintergrundmerkmale und die Zusammenhänge zu den orthografischen Kompetenzen der Schülerinnen und Schüler aus ZuRecht betrachtet. Dabei wird zunächst aufgeschlüsselt, welche Variablen für die Analysen genutzt und anhand welcher Statistiken die Ergebnisse dargestellt werden. Die Befunde beziehen sich sowohl auf die Rechtschreibleistungen der Kinder bezüglich der Schreibung ganzer Wörter als auch auf die Teilkompetenzen der beiden Rechtschreibtests. Sie werden zudem in Beziehung zu Befunden großer Schulleistungsstudien gesetzt, um sie besser einordnen und bewerten zu können. Wenn möglich bzw. vorhanden, wurden dabei insbesondere Studien fokussiert, die ebenfalls die Rechtschreibung von Viertklässlern betrachten. Zum Teil können bei dem Vergleich nur Zusammenhänge auf Ganzwortebene herangezogen werden, da wenige differenzielle Ergebnisse zu Teilkompetenzen vorliegen.

### Weighted-Likelihood-Estimates

Für die Zusammenhangsanalysen werden *Weighted-Likelihood-Estimates* (WLE) nach Thomas A. Warm genutzt. Dabei handelt es sich um Parameterwerte der Personenfähigkeiten, die für jedes Kind herausgegeben werden (Warm, 1989, S. 429 ff.). WLE gehören der Gruppe der Maximum-Likelihood-Schätzer an und werden in der Skalierungssoftware durch ein Marginal-Maximum-Likelihood-Verfahren (MML) geschätzt (Wu et al., 2007, S. 5). Die Schätzwerte sind nicht messfehlerbereinigt, weisen aber den kleinsten Bias unter allen Personenparameterschätzern auf (vgl. dazu die Simulationsstudien von Walter (2005, S. 61 ff.)). Rost bezeichnet die WLE-Methode als „Standardverfahren“ zur Bestimmung der Personenparameter. Zudem sind WLE Rost (2004, S. 315) zufolge die „besten *Punktschätzer* der individuellen Messwerte“. Aus diesen Gründen werden sie für die nachfolgenden Analysen verwendet.

Es wurden sowohl WLE auf Ganzwort- als auch auf Teilkompetenzebene des gutschrift-Tests und SRT berechnet. Analog zu dem Vorgehen in Large-Scale-Assessments, wie z. B. PISA, TIMSS und IGLU bzw. PIRLS, wurden die Logit-Werte auf einen Mittelwert von 500 und eine Standardabweichung von 100 normiert (vgl. z. B. Wendt, Tarelli, Bos, Frey & Vennemann, 2012, S. 61; Frey, Carstensen, Walter, Rönnebeck & Gomolka, 2008, S. 388; Foy et al., 2007, S. 159). Die jeweiligen Relationen der Kinder innerhalb der beiden Tests bleiben aber dennoch auf der Originalmetrik erhalten. Dadurch können die Ergebnisse miteinander verglichen, einfacher veranschaulicht und interpretiert werden (Rauch & Hartig, 2008, S. 241).

### Effektstärke

Um die Mittelwertsunterschiede der Kompetenzwerte zweier Gruppen (die in Form von WLE angegeben werden) statistisch beurteilen zu können, wird im Folgenden die *Ef-*

*fektstärke* oder *-größe* berichtet. Über sie lässt sich die praktische Bedeutsamkeit eines ermittelten Effekts einschätzen (Erdfelder, Faul, Buchner & Cüpper, 2010, S. 363). Wie von Jacob Cohen vorgeschlagen, wird die Effektstärke $d$ bei gleichen Stichprobengrößen aus der Differenz der Mittelwerte $\overline{X}_A$ und $\overline{X}_B$ (mit $\overline{X}_A > \overline{X}_B$) dividiert durch die gemeinsame Standardabweichung $\sigma = \sigma_A = \sigma_B$ berechnet:

$$d = \frac{\overline{X}_A - \overline{X}_B}{\sigma}$$

(in Anlehnung an Cohen, 1988, S. 27)

Unterscheiden sich die Standardabweichungen ($\sigma_A \neq \sigma_B$) in den Stichproben, empfiehlt Cohen, $\sigma$ auf Basis des Mittelwerts der einzelnen Varianzen ($\sigma_A^2$ bzw. $\sigma_B^2$) zu berechnen:

$$\sigma = \sqrt{\frac{\sigma_A^2 + \sigma_B^2}{2}}$$

(in Anlehnung an Cohen, 1988, S. 43 f.)

Sind die Stichprobengrößen für die zu vergleichenden Gruppen nicht gleich ($n_A \neq n_B$), wird die *gepoolte Standardabweichung* verwendet, bei der die beiden zu vergleichenden Gruppenvarianzen vor der Mittelung anhand der Stichprobengrößen gewichtet werden:

$$\sigma = \sqrt{\frac{(n_A - 1)\sigma_A^2 + (n_B - 1)\sigma_B^2}{(n_A - 1) + (n_B - 1)}}$$

(in Anlehnung an Bortz & Döring, 2006, S. 607; Cohen, 1988, S. 67)

Im Kern werden bei der Bestimmung der Effektstärke $d$ also Mittelwertsdifferenzen in Relation zur Varianz der zu vergleichenden Gruppen gesetzt. Dabei entstehen höhere Effekte bei großen Mittelwertsunterschieden und kleinen Varianzen.

Bei der Effektstärke ist keine binäre Wahl – „hat einen Effekt“ oder „hat keinen Effekt“ – entscheidend, sondern das Ausmaß bzw. die Größe des Effekts. Eine Klassifikation der Effektgröße schlug Cohen vor (Erdfelder et al., 2010, S. 363). Sie ist in der Mehrzahl sozial- bzw. humanwissenschaftlicher Forschung etabliert. Danach haben Werte von $d = 0,2$ einen kleinen, $d = 0,5$ einen mittleren und $d = 0,8$ einen großen Effekt (Cohen, 1988, S. 24-27).[10] Cohens Effektgrößenkonvention ist in Tabelle 4.19 mit inhaltlichen Bedeutungsbeispielen aufgeführt.

Von einer Interpretation des Kompetenzunterschieds in Form von Abständen in der Maßeinheit „Schuljahren“ wird abgesehen. Bonsen, Büchter und van Ophysen (2004, S. 204 f.)

[10] Natürlich kann Cohens Einteilung nur als Orientierungsmaß genutzt werden, denn die wirkliche Relevanz von Mittelwertsdifferenzen ist auch immer abhängig von der Thematik bzw. von dem Nutzen oder erzielten Effekt. So können auch kleine Unterschiede, beispielsweise bei explorativen Studien, Hinweise zur Bewertung geben oder im Bereich der Medizin, z. B. im Kontext von lebensrettenden Therapien oder Maßnahmen, eine gesellschaftlich hohe Bedeutsamkeit haben (B. Rasch et al., 2004, S. 67 f.).

| Effektstärke | Klassifikation | Beispiel |
|---|---|---|
| 0,2 | klein | Körpergrößenunterschied bei 15- und 16-jährigen Mädchen |
| 0,5 | mittel | Körpergrößenunterschied bei 14- und 18-jährigen Mädchen |
| 0,8 | groß | Körpergrößenunterschied bei 13- und 18-jährigen Mädchen |

Tabelle 4.19: Klassifikationsbeispiel für die Effektstärke nach Cohen (in Anlehnung an Bortz & Döring, 2006, S. 606, 627)

diskutieren diesen Vergleichsmaßstab und kommen zu dem Ergebnis, dass er willkürlich gesetzt ist und nur sehr eingeschränkte Gültigkeit besitzt: „Tatsächlich lassen sich zu unterschiedlichen Zeitpunkten in Bildungslaufbahnen unterschiedliche Zuwachsraten finden. Zudem unterscheiden sich die Zuwachsraten zwischen Geschlechtern, Klassenstufen, Fächern und Schulformen erheblich. [...] Wenn aber die Zuwächse über die Zeitachse so unterschiedlich sind, macht es keinen Sinn einen universellen Vergleichsmaßstab wie ein Schuljahr anzuwenden. Dieses wäre ein fiktives Schuljahr eines fiktiven Durchschnittsprobanden, für den es kaum empirische Realisierungen gibt."

### Schülerfragebogen in ZuRecht

Die Hintergrundinformationen der Schülerinnen und Schüler aus ZuRecht konnten über einen Kurzfragebogen gewonnen werden, der zusätzlich zu den Rechtschreibtests eingesetzt wurde. Die Kinder wurden gebeten, Angaben zu ihrem Geschlecht, zur weiterführenden Schulform, die sie voraussichtlich im Anschluss an die Grundschule besuchen, und zu der Anzahl der zu Hause verfügbaren Bücher zu machen.[11] Diese Kontextinformationen sollen nun im Zusammenhang mit den orthografischen Kompetenzen der Schülerinnen und Schüler betrachtet werden.[12] Über Balkendiagramme werden sie im Folgenden visualisiert. Dabei sind jeweils als erstes die Ergebnisse auf der Basis korrekt geschriebener Wörter und schließlich die differenzierten Ergebnisse zunächst für die vier Teilkompetenzen des gutschrift-Tests und im Anschluss für die fünf Teilkompetenzen des SRT dargestellt. Als Grundlage wurden die Items aus den im Rahmen der Itemanalyse optimierten Modellen verwendet (vgl. Abschnitte 4.2.1 und 4.2.2). Dies gilt ebenso für die Indikatoren auf Ganzwortebene. Bei dem SRT wurde mit der selektierten Variante gerechnet und es verblieben 76 (von 78) Wörter sowie bei gutschrift 30 (von 35) Wörter im Modell.

[11] Die Erfassung weiterer Angaben, wie beispielsweise des Migrationshintergrundes oder der Sprachgewohnheiten, waren aufgrund der geforderten Kürze des Schülerfragebogens leider nicht möglich.

[12] Hiermit kann natürlich nur ein minimaler Ausschnitt von Bedingungsfaktoren schulischer Leistung erhoben werden. In Large-Scale-Assessments wird die Verknüpfung von Schülerleistungen und deren Bedingungen in einem Rahmenmodell formuliert und analysiert (vgl. Hornberg, Bos, Buddeberg, Potthoff & Stubbe, 2007, S. 22). Dabei werden vielfältige Instrumente zur Erfassung der familiären, schulischen und außerschulischen Kontextfaktoren sowie individueller Merkmale und Voraussetzungen der Kinder genutzt, um belastbare Aussagen formulieren zu können.

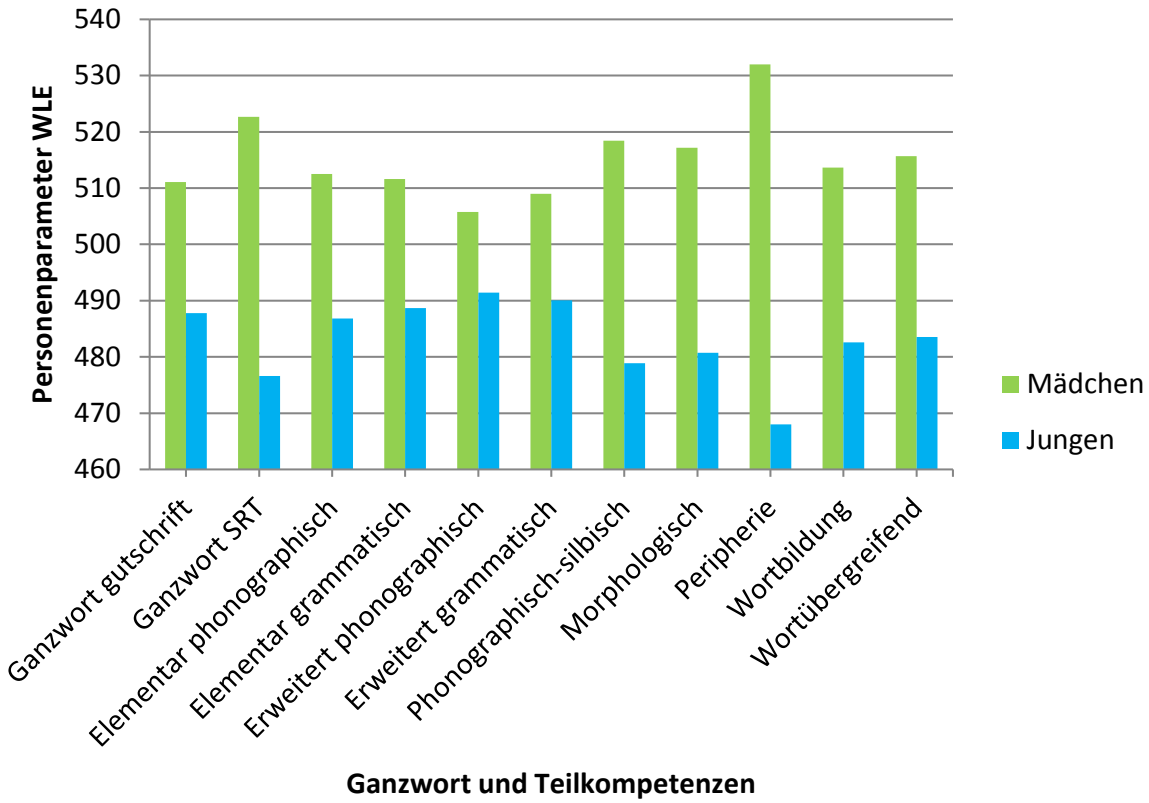



Abbildung 4.7: Zusammenhang zwischen Rechtschreibkompetenz und Geschlecht

Die Mittelwertsunterschiede in den Testleistungen werden, wie zuvor beschrieben, zusätzlich über die Angabe von Effektstärken berichtet. Die Datenbasis bilden die 547 Kinder aus dem gemeinsamen neundimensionalen Skalierungslauf der beiden Rechtschreibtests (vgl. Abschnitt 4.2.3). Da nicht immer zu allen dieser Schülerinnen und Schüler vollständige Kontextinformationen vorliegen, fällt die Stichprobengröße bei den einzelnen Analysen kleiner aus, liegt aber stets über 500. Angesichts dieses Stichprobenumfangs wird auf den explorativen Charakter der Analysen verwiesen.

### 4.3.1 Geschlecht

Von den 547 Schülerinnen und Schülern, die sowohl den gutschrift-Test als auch den SRT bearbeitet haben, liegen von insgesamt 510 auch Angaben zum Geschlecht vor. Die Mädchen und Jungen teilen sich relativ gleichmäßig auf die Stichprobe auf: es haben 248 Mädchen und 262 Jungen die beiden Tests bearbeitet. In Abbildung 4.7 sind die Fähigkeitswerte basierend auf den WLE-Schätzern getrennt nach Geschlecht wiedergegeben.

Sowohl auf Ganzwortebene als auch innerhalb aller Teilkompetenzen des gutschrift-Tests sowie des SRT haben Mädchen einen Vorsprung gegenüber Jungen. Abbildung 4.7 zeigt eine asymmetrische Kompetenzverteilung der Geschlechter. In beiden Rechtschreibtests liegen die Leistungen der Schülerinnen jeweils über dem Durchschnitt und die der Schüler

unterhalb von 500 Punkten. Beim SRT sind im Vergleich zum gutschrift-Test insgesamt größere Mittelwertsunterschiede zu beobachten. Die berechnete Effektstärke beträgt für den SRT auf der Ebene ganzer Wörter $d = 0,47$, was nach der Klassifikation von Cohen einem mittleren Effekt nahe kommt. In gutschrift findet sich ein unbedeutender Effekt von $d = 0,23$.

Die testübergreifend stärkste Differenz auf Teilkompetenzebene zwischen den Geschlechtern weist der Peripheriebereich mit 64 Punkten auf, was mit $d = 0,67$ als mittlerer bis großer Effekt zu werten ist. Der zweithöchsten Effektstärke innerhalb der Teilkompetenzen von $d = 0,40$ entsprechen die durchschnittlichen Leistungsunterschiede mit rund 40 Punkten im phonographisch-silbischen Bereich. Bei den weiteren SRT-Teilkompetenzen liegen die gemittelten Testleistungen der Jungen etwa 33 Punkte unter denen der Mädchen. In gutschrift sind die Leistungen zwischen Schülerinnen und Schülern in den erweiterten Kompetenzdimensionen homogener als in den elementaren Teilkompetenzen: sie liegen bei 14 (erweitert phonographisch) bzw. 19 (erweitert grammatisch) im Vergleich zu 26 (elementar phonographisch) und 23 (elementar grammatisch) Differenzpunkten. Es lässt sich festhalten, dass es bei der Beherrschung der Teilkompetenzen im gutschrift-Test und SRT zwischen den Geschlechtern gleich gerichtete Unterscheide gibt, wobei diese im Peripheriebereich bedeutsam sind.

Die hier berichteten Ergebnisse decken sich mit den Befunden von Large-Scale-Assessments, in denen ebenfalls die beiden Tests in Klasse 4 eingesetzt worden sind. Für gutschrift wird auf Ganzwortebene in IGLU-E 2006 und KESS 4 ein sehr ähnlicher Leistungsvorsprung der Mädchen von jeweils $d = 0,20$ dokumentiert (Kowalski et al., 2010, S. 36; May, 2006b, S. 122). Für die Teilkompetenzen liegen hier keine Werte vor. Im vorangegangenen IGLU-Erhebungszeitpunkt (2001) sind Ergebnisse zu diesen über die Angabe von Prozent-korrekt-Werten veröffentlicht. Die Schülerinnen schneiden hier ebenfalls besser ab, wenn auch nur knapp: die Differenzen betragen für die erweiterte grammatische Teilkompetenz 3,7 Prozentpunkte und für die weiteren Teilkompetenzen rund 1 Prozentpunkt. Auf Gesamtwortebene erhöht sich der Vorsprung in IGLU 2001 auf 6 Prozentpunkte (Valtin, Badel et al., 2003, S. 250). Für den SRT zeigen sich übereinstimmende Ergebnisse mit den Befunden der Voruntersuchung zu IGLU 2006. Auch hier nimmt der Peripheriebereich eine Sonderstellung ein. In der Pilotstudie wurden ca. 70 Punkte (Werte ebenfalls normalisiert auf M = 500 und SD = 100) im Peripheriebereich und in den weiteren Prinzipien 30 Punkte mehr zugunsten der Mädchen ermittelt (Voss et al., 2007, S. 28).

Für die HSP, die im Rahmen von KESS eingesetzt worden ist, liegen differenzierte Ergebnisse vor. Auf der Ebene ganzer Wörter wird ein Effekt von $d = 0,25$ ermittelt und von $d = 0,22$ für den Bereich der Graphemtreffer (May, 2006b, S. 122). Diese Werte fallen ebenfalls ähnlich zu den hier ermittelten Werten bei der Ganzwortebene von gutschrift aus. Im Rahmen der unterschiedenen HSP-Strategien sind Effektstärken von $d = 0,17$ für die alphabetische, $d = 0,25$ für die orthografische, $d = 0,11$ für die morphematische Strategie, $d = -0,06$ für den Bereich überflüssiger orthografischer Elemente sowie $d = -0,31$ für

den Bereich der Oberzeichenfehler dokumentiert (May, 2006b, S. 122).[13] Das Ergebnis zu den Oberzeichenfehlern wird von May (2006b, S. 122) so interpretiert, dass Mädchen im Vergleich zu Jungen ihre Schreibprodukte stärker überprüfen. Cohens Klassifikation zufolge sind die Mittelwertsdifferenzen vernachlässigbar, fallen aber vollständig zugunsten der Mädchen aus.

Die IQB-Pilotstudie zu den Bildungsstandards im Primarbereich dokumentiert auf der Ebene ganzer Wörter einen Vorteil der Mädchen in Höhe von $d = 0,22$ (Böhme & Bremerich-Vos, 2009, S. 346). Auf Ganzwortebene kann damit der hier für den gutschrift-Test ermittelte Leistungsvorsprung der Mädchen über unterschiedliche Studien (IGLU-E 2006, KESS 4, IQB-Pilotstudie zu den Bildungsstandards) reproduziert werden. Beim IQB-Ländervergleich in Klasse 4 fällt die Differenz mit $d = 0,33$ hingegen höher aus (Böhme & Roppelt, 2012, S. 181).[14] Das Geschlecht wird hier in Zusammenhang mit einem globalen Rechtschreibwert gesetzt, der sich auf der Grundlage eines eindimensionalen Skalierungslaufes mit Lupenstellen, die innerhalb der Wörter betrachtet werden, ergibt (vgl. Abschnitte 2.1.2 und 2.1.3). Die Effektstärke im IQB-Ländervergleich ist im Vergleich zu den weiteren betrachteten Domänen die höchste: Die Differenzen betragen für die Bereiche Lesen, Zuhören und Mathematik (global) in der vierten Klasse $d = 0,24$, $d = 0,03$ und $d = -0,16$[15] (Böhme & Roppelt, 2012, S. 181).

Die Analysen zu den Genderdifferenzen in sprachlichen Bereichen sind mit Blick auf die nationalen und internationalen Befunde von z. B. PIRLS/IGLU und den IQB-Ländervergleich nicht unerwartet. Die Domäne des Lesens stellt hier einen bereits ausführlich dokumentierten Bereich dar, in der die Ergebnisse ebenfalls und wiederholt zugunsten der Mädchen ausfallen (vgl. z. B. Valtin, Bos, Buddeberg, Goy & Potthoff, 2008, S. 77 f.; Drechsel & Artelt, 2007, S. 234). In Klasse 4 wird im Rahmen von IGLU 2006 ein durchschnittlicher Kompetenzvorsprung der Schülerinnen von 7 Punkten (Hornberg, Valtin, Potthoff, Schwippert & Schulz-Zander, 2007, S. 202), in IGLU 2011 von 8 Punkten (Bos, Bremerich-Vos, Tarelli & Valtin, 2012, S. 127) und bei den IQB-Bildungsstandards von 24 Punkten sowie einer Effektstärke von $d = 0,24$ dokumentiert (Böhme & Roppelt, 2012, S. 181).

### 4.3.2 Weiterführende Schulform

In dem Fragebogen wurden die Kinder gefragt, auf welche weiterführende Schulform sie nach den Sommerferien voraussichtlich wechseln. Die Schulformwahl im Anschluss an die Primarstufe richtete sich in NRW zum Zeitpunkt der Erhebung nach den Empfehlungen im Grundschulgutachten. Falls die Auffassungen der Grundschule und Eltern divergierten, wurde das Schulamt hinzugezogen, das die Eignung über einen Prognoseunterricht prüfte

[13] Die Effektstärken der überflüssigen orthografischen Elemente und der Oberzeichenfehler sind aufgrund der Polung negativ; über sie werden Differenzen im Mittelwert der Fehler – und eben nicht der Richtigschreibungen – berechnet.

[14] Weitere Analysen mit Hintergrundmerkmalen liegen in dem Berichtsband für die Orthografie nicht vor.

[15] Hier kann an dem Vorzeichen abgelesen werden, welches Geschlecht besser abschneidet; es ist negativ, wenn die Jungen bessere Kompetenzwerte aufweisen.

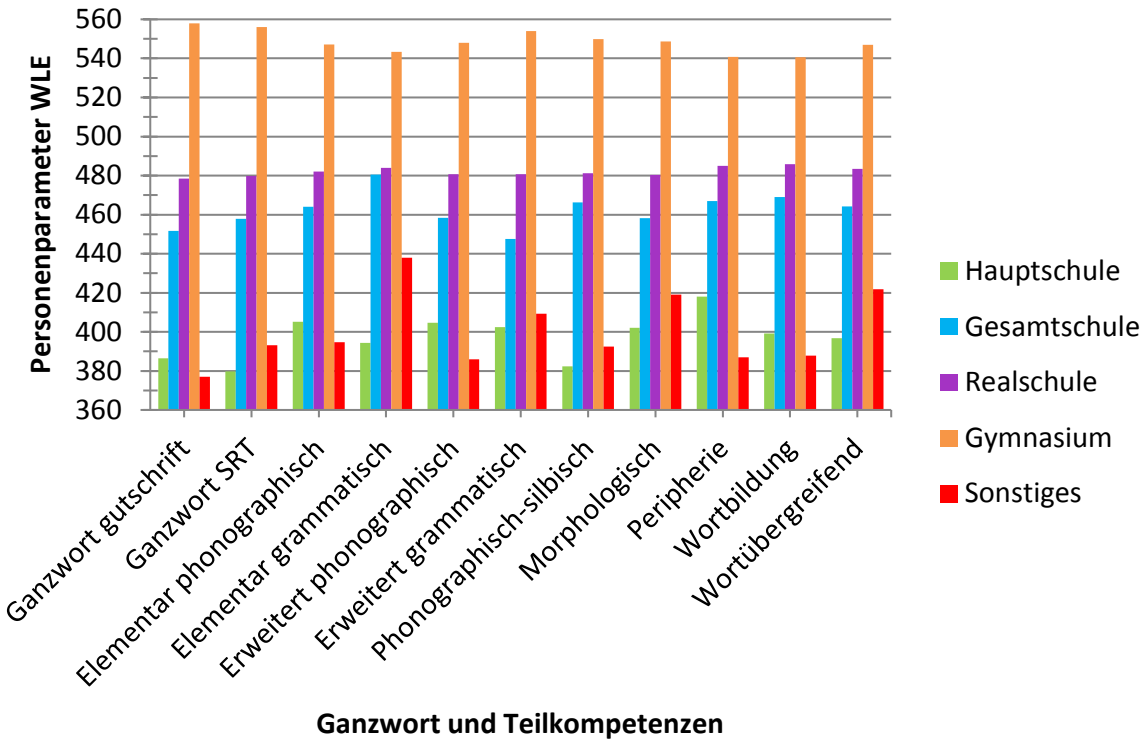



Abbildung 4.8: Zusammenhang zwischen Rechtschreibkompetenz und weiterführender Schulform

(Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen, 2008, S. 4, § 11(4)). Die frühe Trennung der Schülerinnen und Schüler nach Klasse 4[16] auf eine weiterführende allgemeinbildende Schule ist in Deutschland eine Besonderheit des gegliederten Schulsystems (van Ackeren & Klemm, 2011, S. 49 ff.). Als Antwortmöglichkeiten standen die zum Zeitpunkt der Erhebung möglichen staatlichen Schulformen Haupt-, Gesamt- und Realschule sowie Gymnasium und Förderschule zur Verfügung. Diese Auswahl wurde durch eine „weiß nicht"-Kategorie ergänzt. Ebenfalls wurde die Grundschule als Kategorie eingefügt, um die Möglichkeit der Klassenwiederholung zu berücksichtigen.

In Abbildung 4.8 ist das Abschneiden der Schülerinnen und Schüler in den beiden Rechtschreibtests getrennt nach weiterführender Schulform dargestellt. Zur Schulformfrage liegen 520 Antworten vor. Den kommenden Wechsel auf die Hauptschule gedenken 41 Kinder, auf die Gesamtschule 84 Kinder, auf die Realschule 180 Kinder und auf das Gymnasium die Mehrheit von 210 Kindern anzutreten. Fünf Schülerinnen und Schüler kreuzten die Antwortmöglichkeiten Förder-, Grundschule oder „weiß nicht" an. Diese Antworten wurden zu der Kategorie „Sonstiges" zusammengefasst. Sie werden aufgrund der kleinen Anzahl nicht weiter berücksichtigt, sind aber der Vollständigkeit halber mit in Abbildung 4.8 aufgeführt.

[16]Eine Ausnahme bilden nur wenige Bundesländer mit einer sechsjährigen Grundschulzeit.

Die mittleren Leistungswerte der voraussichtlich zukünftigen Gymnasiasten fallen auf Ganzwort- und Teilkompetenzebene, wie Abbildung 4.8 zeigt, für beide Tests mit rund 550 Punkten überdurchschnittlich aus. In allen weiteren Schulformen liegen die Testergebnisse durchgängig unterhalb des Durchschnitts. Das Balkendiagramm zeigt klar eine Rangreihenfolge an. So kann eine streng monotone Steigung der durchschnittlichen Kompetenzwerte von der Hauptschule über die Gesamtschule und Realschule bis zum Gymnasium hinweg beobachtet werden. Auf Basis dieser stufenweisen Reihenfolge sind die größten Leistungssprünge für die beiden Tests zwischen der Haupt- und Gesamtschule sowie zwischen der Realschule und dem Gymnasium vorhanden, während sie für die Gesamt- und die Realschule vergleichsweise gering ausfallen. Die Leistungsdifferenzen zwischen der Haupt- und Gesamtschule sowie zwischen der Realschule und dem Gymnasium betragen auf Gesamtwortebene für beide Tests 65 bis 79 Punkte. Die dazugehörigen Effektstärken liegen zwischen $d = 0,76$ und $d = 0,96$. Für die einzelnen Teilkompetenzen sind sie ähnlich hoch. Innerhalb des phonographisch-silbischen Prinzips werden diese zwischen den Kategorien Hauptschule und Gesamtschule mit $d = 1,00$ sogar noch überstiegen.

Die maximalen Leistungsabstände über alle Schulformen hinweg, also zwischen der Hauptschule und dem Gymnasium, betragen für den Bereich richtig geschriebener Wörter über 170 Punkte für beide Tests (vgl. Abbildung 4.8). Bei gutschrift entspricht das einem Effektmaß von $d = 1,68$ und bei dem SRT von $d = 2,02$. Diese und die zuvor berichteten standardisierten Mittelwertsdifferenzen demonstrieren einen bedeutenden Leistungsvorsprung bzw. -rückstand der Schülerinnen und Schüler im Bereich der Orthografie am Ende der Grundschulzeit.

Die streng monotone Steigung der Kompetenzwerte in Abbildung 4.8 zeigt einen systematischen Zusammenhang zwischen der Schulform und der Rechtschreibung. Die Befunde der ersten IGLU-Untersuchung verdeutlichen diesen Zusammenhang ebenfalls: Als Übergangsempfehlung bekommen 75 Prozent der Schülerinnen und Schüler auf der niedrigsten orthografischen Kompetenzstufe die Hauptschule und 76 Prozent der Kinder auf der höchsten orthografischen Kompetenzstufe das Gymnasium ausgesprochen (Valtin, Badel et al., 2003, S. 248). In DESI werden Leistungsunterschiede nach Schulform für die Neuntklässler ausgewiesen. Hier zeigt sich, dass am Ende der Sekundarstufe I über die Hälfte der Gymnasiasten die obersten beiden Kompetenzniveaus in Rechtschreiben erreichen, während etwa ein Viertel bis ein Drittel der Hauptschüler die beiden untersten Niveaus nicht überschreiten (G. Thomé & Eichler, 2008, S. 110). Beck et al. (2009, S. 45) kommen sogar zu dem Schluss, dass die Rechtschreibung eine „größere Auswirkung“ auf den Besuch der weiterführenden Schulform hat, als das Lesen: „Auffallend (aber nicht erwartungswidrig) ist der signifikant hohe Anteil des Bildungsgangs Hauptschule in der Gruppe der Rechtschreibschwachen (60 Prozent) im Vergleich zur Gruppe der Leseschwachen (40 Prozent). Das bedeutet, dass von den 10 Prozent der schwächsten Rechtschreiber 60 Prozent die Hauptschule besuchen. [...] Von den Schülern mit großen Leseproblemen besuchen nur 40 Prozent diesen Schulzweig.“

Insgesamt lässt sich nicht eindeutig eine Teilkompetenz identifizieren, die einen maßgeblichen Einfluss auf den Besuch einer Schulform hat. Dieses Resultat demonstrieren

auch die differenzierten Untersuchungen zur Kompetenzverteilung für die Domäne Naturwissenschaften im Rahmen von PISA. Es finden sich keine schulformcharakteristischen Profilläufe über die dort untersuchten sieben naturwissenschaftlichen Teilkompetenzen. Stattdessen wird von einer „Gleichförmigkeit der Profilverläufe“ gesprochen (Rost et al., 2004, S. 137).

### 4.3.3 Bildungshintergrund

Analog zu IGLU wird die Anzahl der Bücher im Haushalt als verlässlicher Indikator für den Bildungshintergrund des Kindes angesehen: „Der Buchbesitz ist ein Merkmal der Bildungsnähe der Elternhäuser – die Bildungsnähe wiederum ist ein Hinweis auf die Sozialschicht. Die Erklärungskraft dieser Variable für die schulischen Kompetenzen von Kindern hat sich schon in zahlreichen internationalen Studien gezeigt.“ (Bos, Schwippert & Stubbe, 2007, S. 228) Damit ist es möglich, auf soziale Ausgangssituationen von Kindern zu schließen und Zusammenhänge zu Kompetenzen zu berechnen. Natürlich bieten sich noch eine Vielzahl weiterer Indizes an. Im Rahmen dieser Erhebung konnten aber keine weiteren Fragen zur sozialen Herkunft gestellt werden. Die Kategorien der Variable „Bücheranzahl“ sind in dem Fragebogen von ZuRecht identisch mit der IGLU-Hauptuntersuchung und lauten (Bos, Schwippert & Stubbe, 2007, S. 229):

1. 0-10 Bücher
2. 11-25 Bücher
3. 26-100 Bücher
4. 101-200 Bücher
5. über 200 Bücher

Die Anzahl wurde in dem Kurzfagebogen durch Bilder mit Büchern in Regalen visualisiert. Zudem erhielten die Kinder den Hinweis, Zeitschriften, Zeitungen und Schulbücher nicht mitzuzählen. Abbildung 4.9 schlüsselt die Verteilung der orthografischen Kompetenzwerte nach der im Haushalt verfügbaren Anzahl an Büchern auf. Für die Analysen mussten 38 Fälle ausgeschlossen werden, für die keine Antwort auf diese Frage vorliegt. Insgesamt basieren die Ergebnisse damit auf 509 Schülerinnen und Schülern.

Die Kompetenzwerte steigen auf Basis vollständig richtiger Wortschreibungen beim gutschrift-Test und dem SRT im Verhältnis zur Bücheranzahl monoton an. Die Leistungsspanne zwischen „bis 10“ und „über 200 Bücher“ ist dementsprechend am größten: Für gutschrift beträgt sie 35 Punkte bzw. $d = 0,34$ und für den SRT 80 Punkte und hat einen mittleren bis hohen Effekt von $d = 0,74$. Auf Teilkompetenzebene kann die höchste Differenz und Effektstärke von 81 Punkten bzw. $d = 0,80$ zwischen Kategorie 1 und 4 im Peripheriebereich ausgemacht werden.

Die Abstände der mittleren Kompetenzwerte sind zwischen den Bücherkategorien auf Wort- und Teilkompetenzebene ungleichmäßig. Abbildung 4.9 zeigt, dass insbesondere große Kompetenzunterschiede zwischen wenigen Büchern (0-10 Bücher), einer mittleren

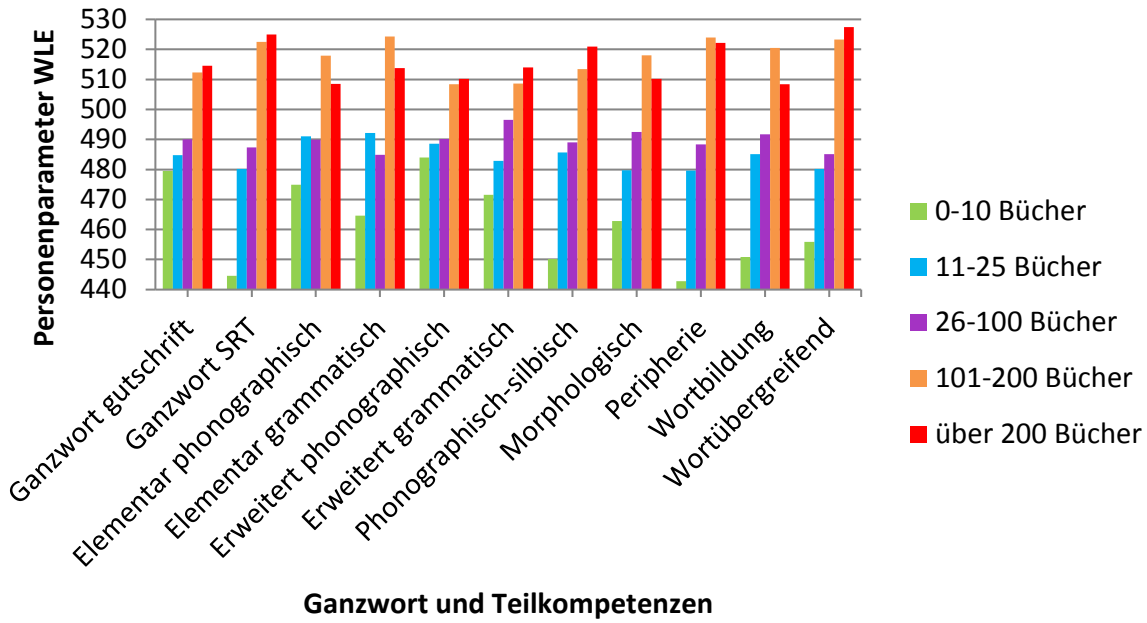



Abbildung 4.9: Zusammenhang zwischen Rechtschreibkompetenz und Bücheranzahl

Anzahl an verfügbaren Büchern (11-100 Bücher) sowie vielen Büchern (ab 101 Bücher) erkennbar sind. Kleinere Mittelwertsdifferenzen sind dementsprechend bei den Antwortmöglichkeiten zwei und drei sowie vier und fünf vorhanden. Bei dem gutschrift-Test fallen die Leistungswerte zwischen diesen Kategorien in den elementaren Kompetenzen sogar leicht ab. Für den SRT trifft diese Beobachtung auf die vierte und fünfte Antwortmöglichkeit (bei dem morphologischen Prinzip, dem Peripheriebereich und dem Prinzip der Wortbildung) zu.

Die Kompetenzwerte aller Schülerinnen und Schüler liegen ab „über 100 Bücher“ jeweils oberhalb des Mittelwerts von 500 Punkten. Ab dieser Kategorie finden die beschriebenen mehrfach auftretenden Reihenfolgewechsel in beiden Tests im Kompetenzwert statt. Der Besitz von mehr als 100 Büchern kann damit als eine wichtige Schwelle angesehen werden, ab dem die Testleistung überdurchschnittlich ausfällt, wobei über dieser Schwelle das Mehr an Büchern keinen so starken Einfluss auf die Testleistung zu haben scheint.

Bezüglich des orthografischen Kompetenzstandes kann eine Kopplung von Bildungsbenachteiligung und sozialer Herkunft beobachtet werden. Es zeigt sich das erwartungskonforme Ergebnis, dass Schülerinnen und Schüler bei einer Vielzahl von Büchern bessere Testleistungen erbringen. Der Zusammenhang zwischen Rechtschreibkompetenz und Bildungshintergrund zeigte sich u. a. bereits im ersten Erhebungszyklus von IGLU. Hier liegen ausschließlich prozentuale Ergebnisse auf Wortebene für die Gruppe der Kinder des unteren und oberen Quartils (gebildet nach der Anzahl an Einzelfehlern) vor. Diese bestätigen ebenfalls die bessere Testleistung bei Kindern mit einem Zugang zur Welt des Lesens. Die Gruppenanteile verteilen sich wie folgt: 39 Prozent der leistungsstarken Rechtschreiber (oberes Quartil) und 15,5 Prozent der leistungsschwachen Schreiber (unteres Quartil) sind in einem Haushalt mit mehr als 200 Büchern vertreten. Aus Familien mit

weniger als 25 Büchern stammen 9,7 Prozent der leistungsstarken Rechtschreiber und 30,7 Prozent der leistungsschwachen Schreiber (Valtin, Badel et al., 2003, S. 254). Für KESS liegen ebenfalls keine differenziellen Ergebnisse für die Subskalen der HSP vor, aber ein globaler Wert, der angibt, dass die Kompetenzleistung bei über 100 Büchern höher – in etwa 15 Punkte (bei M = 100 und SD = 30) – liegt, als bei unter 100 Büchern (Pietsch, 2007, S. 14 f.). In der IGLU-Voruntersuchung sind zum Teil differenzierte Befunde dokumentiert. Diese weisen durchschnittliche Testleitungen auf Wortebene von 560 Punkten bei Familien aus, die angeben, mehr als 100 Kinderbücher zu besitzen. Bei maximal 10 Kinderbüchern erreichen die Schülerinnen und Schüler weniger Kompetenzpunkte: für die beiden Kernbereiche sowie die Großschreibung rund 400 und für den Peripheriebereich 445 Punkte (Voss et al., 2007, S. 28).

Die berichteten Ergebnisse für die Orthografie zeichnen das Bild von Schulleistungsstudien und den dort erhobenen sprachlichen Domänen, wie z. B. dem Lesen, gut nach. So dokumentieren Bos, Schwippert und Stubbe (2007, S. 233) einen Unterschied in der Lesekompetenz von 40 Punkten zugunsten der Kinder, bei denen zu Hause mehr als 100 Bücher verfügbar sind, im Vergleich zu den Kindern mit weniger als 100 Büchern.

### 4.3.4 Zusammenschau

In diesem Kapitel wurden die orthografischen Testleitungen der Schülerinnen und Schüler aus ZuRecht in Beziehung zu den drei erhobenen Hintergrundvariablen Geschlecht, weiterführende Schulform und Bücheranzahl gesetzt. Die Testleistungen wurden zu diesem Zweck auf eine vergleichbare Skala transformiert und die Berechnung von Effektstärken nach Cohen sowie deren Größenkonventionen und Vorteil dargestellt. Zudem wurden die Analyseergebnisse in einem Gesamtzusammenhang betrachtet, indem Ergebnisse großer Schulleistungsstudien zur Rechtschreibung und zum Teil zu weiteren Domänen hinzugezogen wurden. Die Personenwerte wurden vergleichend für die beiden Rechtschreibtests gutschrift und SRT in Balkendiagrammen dargestellt und erläutert. Der testübergreifende Vergleich erfolgte auf Ganzwortebene und differenziert nach Teilkompetenzen.

Im Bereich der Genderanalysen sind insbesondere die Ergebnisse des Peripheriebereichs des SRT hervorzuheben. Hier wurde ein mittlerer bis großer Effekt bestimmt. Mädchen schneiden im Peripheriebereich damit deutlich besser ab als Jungen. Ausnahme- und Fremdwortschreibungen beherrschen sie den Befunden zufolge besser. Eine mögliche Interpretation wäre der Einfluss des Lesens. Neben den besseren Leseleistungen könnten Schülerinnen durch ihr erhöhtes außerschulisches Leseverhalten (Bos, Bremerich-Vos et al., 2012, S. 127; Hornberg, Valtin et al., 2007, S. 210 ff.) einen größeren Wortschatz besitzen und damit auch mehr fremde Wörter bzw. Wortschreibungen aufnehmen als im Unterricht gelehrt werden. Es könnte auch ein Zusammenhang zu der Beherrschung der weiteren Teilkompetenzen bestehen. Eventuell sind diese Voraussetzungen für die Schreibungen der Struktureinheiten des Peripheriebereichs.

In allen weiteren SRT- und gutschrift-Teilkompetenzen und für die Schreibung insgesamt richtiger Wörter zeigen Schülerinnen ebenfalls schriftsprachliche Vorteile gegenüber Schü-

lern. Für den SRT sind die Vorsprünge der Mädchen dabei noch stärker ausgeprägt, was beispielsweise dem Geschichtsinhalt des Diktats, in dem es u. a. um Pferde geht, geschuldet sein könnte. Hier wären zukünftige weiterführende Analysen zum Schreibverhalten, bzw. nach typischen Mädchen- bzw. Jungenwörtern mittels *differential item functioning* (vgl. Embretson & Reise, 2000, S. 249 ff.), aufschlussreich.

Die Schulformanalysen zeigen eine große Leistungsheterogenität mit bedeutenden Mittelwertsdifferenzen von Kindern am Ende der Grundschulzeit an. Der Erwerb von Rechtschreibkompetenz ist zu Beginn der Sekundarstufe nicht abgeschlossen und muss weiter ausgebaut und gefördert werden, was z. B. insbesondere für die Hauptschulen wichtig erscheint. Die Mittelwertsunterschiede sind zwischen dieser Schulform und dem Gymnasium am höchsten. Bei allen Teilkompetenzen und auf Wortebene weisen die Gymnasiasten weit überdurchschnittliche Kompetenzwerte auf. Gefolgt werden diese von den Real- und Gesamtschülern, die über ein ähnliches Kompetenzniveau verfügen, das knapp unterhalb des Mittelwertes liegt. Die Hauptschüler bilden die Gruppe mit den schlechtesten Ergebnissen und sind mit Werten von rund 400 Punkten weit abgeschlagen von den übrigen Kindern der Stichprobe.

Bei der Erhebung des Bildungshintergrundes mittels der Erfragung der Bücheranzahl wurde ein Zusammenhang zur Rechtschreibkompetenz beschrieben, der zwar nicht so deutlich wie für die Schulformanalysen ausfiel, aber dennoch mittlere bis große Effektstärken aufwies. Entscheidend für eine überdurchschnittlich ausgeprägte orthografische Kompetenz scheint ein Besitz über 100 Büchern zu sein; gleichzeitig verringert sich ab dieser Menge aber auch die Deutlichkeit des Einflusses auf die Testleitungen.

## 4.4 Profilanalysen

In Abschnitt 4.4.1 werden die Leistungsergebnisse ausgewählter Schülerinnen und Schüler über die Angabe der Richtigschreibungen auf Wort- und Teilkompetenzebene beschrieben. Hierbei werden intraindividuelle Leistungsunterschiede in den Teilkompetenzen der Kinder explorativ analysiert. Während also in 4.4.1 einzelne Profile von Kindern dargestellt und interpretiert werden, erfolgt hingegen in Abschnitt 4.4.2 eine Analyse über die gesamte Stichprobe. Es werden latente Profilanalysen berechnet, um Schülerinnen und Schüler mit gemeinsamen Kompetenzprofilen zu identifizieren. Die Ergebnisse werden dabei anhand der in Abschnitt 3.3.3 eingeführten statistischen Maßzahlen beurteilt.

### 4.4.1 Einzelfalldarstellung und -interpretation

Zu berücksichtigen gilt bei der Betrachtung von einzelnen Diktaten, dass beide Rechtschreibtests komprimierte Informationen eines Messzeitpunktes zur Verfügung stellen. Auf Ebene der Diagnose und Förderung sind genauere und umfangreichere Betrachtungen der Schreibprodukte unverzichtbar. Leistungsstudien eignen sich daher nicht für eine Individualdiagnostik (Tarelli, Wendt, Bos & Zylowski, 2012, S. 56). In diesem Rahmen

werden daher ausschließlich qualitativ und explorativ einzelne Schülerinnen und Schüler betrachtet. Es soll geprüft werden, ob es beispielsweise Kinder gibt, die testübergreifend spezifische Teilkompetenzen besser beherrschen als andere Schülerinnen und Schüler, oder ob Kinder allgemein in allen Teilkompetenzen vergleichbare Leistungen aufweisen. Die Auswahl der hier im Folgenden dargestellten Fallbeispiele beruht auf einer Ziehung, die sich im Rahmen des Dateneingabeprozesses ergeben hat, bei dem interessante Fälle notiert wurden.

In Tabelle 4.20 ist das Abschneiden der ausgewählten Schülerinnen und Schüler auf Ganzwortebene und in den vier Teilkompetenzen des gutschrift-Tests sowie den fünf Teilkompetenzen des SRT dargestellt. Als Maßzahlen werden die absolute Anzahl und der prozentuale Anteil der insgesamt richtig geschriebenen Wörter und Analyseeinheiten angegeben. Als Referenzwerte sind die Ergebnisse für die gemeinsame Stichprobe der beiden Tests ($n$ = 547) in der Spalte „Gesamt“ angegeben. Zusätzlich enthält die Spalte „Max.“ die Anzahl der maximal richtig zu schreibenden Struktureinheiten. Die Werte basieren jeweils auf den aus der Itemanalyse resultierenden, optimierten Modellen. Dies gilt nicht nur für die Teilkompetenzen, sondern auch für die Schreibung ganzer Wörter, weshalb bei gutschrift 30 und bei dem SRT 76 maximal richtige Wörter möglich sind.

### Fallbeispiel 1

Bei Fallbeispiel 1 handelt es sich um ein Kind, das auf der Ebene richtig geschriebener Wörter bei dem gutschrift-Test im Verhältnis zur gesamten Stichprobe unterdurchschnittliche Ergebnisse erzielt, da es nur die Hälfte der Diktatwörter korrekt verschriftlicht. Beim SRT liegen die Leistungswerte auf mittlerem Niveau. Differenzierter ist die Betrachtung der Teilkompetenzen, denn hier fallen die vergleichsweise niedrigen Werte in den erweiterten gutschrift-Teilkompetenzen sowie dem Peripheriebereich auf. Die Leistungen liegen in den erweiterten Kompetenzen bei 55 und 68 Prozent sowie dem Peripheriebereich bei 52 Prozent, während in allen weiteren Teilkompetenzen zwischen 83 und 100 Prozent der Analyseeinheiten richtig geschrieben werden. Im Vergleich mit der Gesamtstichprobe schneidet das Kind in der erweiterten phonographischen Kompetenz und dem Peripheriebereich 22 Prozentpunkte sowie in der erweiterten grammatischen Kompetenz 3 Prozentpunkte schlechter ab, wobei es ansonsten eher mittlere bis überdurchschnittliche Leistungen erzielt.

Zur Charakterisierung des Kindes werden zudem die enthaltenen Informationen aus dem Fragebogen ergänzt. Die Auswertung der Kontextdaten zeigt, dass es sich um ein Mädchen handelt, das vermutlich im Anschluss an die Grundschule auf die Realschule gewechselt hat. Zudem gab die Schülerin an, dass in ihrem Zuhause 101 bis 200 Bücher zur Verfügung stehen. Dies spricht für ein vermutlich bildungsnahes Elternhaus.

| Wortebene/ Teilkompetenzen | Max. | Gesamt | Fallbeispiele 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|
| gutschrift Ganzwort | 30 | 18 | 15 | 27 | 2 | 6 | 16 |
| | 100% | 60% | 50% | 90% | 7% | 20% | 53% |
| SRT Ganzwort selektiert | 76 | 49 | 49 | 60 | 13 | 8 | 34 |
| | 100% | 65% | 64% | 79% | 17% | 11% | 45% |
| Elementar phonographisch | 23 | 20 | 22 | 23 | 17 | 17 | 19 |
| | 100% | 88% | 96% | 100% | 74% | 74% | 83% |
| Elementar grammatisch | 15 | 13 | 15 | 14 | 10 | 8 | 14 |
| | 100% | 87% | 100% | 93% | 67% | 53% | 93% |
| Erweitert phonographisch | 22 | 17 | 12 | 22 | 1 | 8 | 16 |
| | 100% | 77% | 55% | 100% | 5% | 36% | 73% |
| Erweitert grammatisch | 22 | 16 | 15 | 22 | 8 | 8 | 17 |
| | 100% | 71% | 68% | 100% | 36% | 36% | 77% |
| Phonographisch-silbisches | 80 | 73 | 74 | 78 | 56 | 25 | 71 |
| Prinzip im Kernbereich | 100% | 91% | 93% | 98% | 70% | 31% | 89% |
| Morphologisches Prinzip | 52 | 43 | 45 | 50 | 25 | 8 | 38 |
| im Kernbereich | 100% | 82% | 87% | 96% | 48% | 15% | 73% |
| Peripheriebereich | 23 | 17 | 12 | 17 | 7 | 8 | 16 |
| | 100% | 74% | 52% | 74% | 30% | 35% | 70% |
| Prinzip der Wortbildung | 25 | 22 | 21 | 24 | 18 | 4 | 21 |
| | 100% | 87% | 84% | 96% | 72% | 16% | 84% |
| Wortübergreifendes | 29 | 24 | 24 | 29 | 14 | 10 | 9 |
| Prinzip | 100% | 84% | 83% | 100% | 48% | 34% | 31% |

Tabelle 4.20: Leistungswerte ausgewählter Kinder in den Kompetenzmodellen

### Fallbeispiel 2

Fallbeispiel 2 ist ein Kind mit überdurchschnittlichen Ergebnissen in allen Subskalen. Es erreicht in acht Teilkompetenzen einen Wert von über 90 Prozent, von denen vier sogar einen Wert von 100 Prozent annehmen. Demnach werden die Indikatoren der beiden phonographischen und der erweiterten grammatischen Kompetenzen sowie des wortübergreifenden Prinzips fehlerfrei verschriftet. Ausschließlich der Peripheriebereich bereitet dem Kind noch Probleme. Der Wert von 74 Prozent entspricht zwar dem Durchschnitt der Gesamtstichprobe, liegt aber in Relation unter den restlichen Kompetenzen. Wie in Abschnitt 2.4 gezeigt, umfasst der Peripheriebereich Struktureinheiten, wie das Dehnungs-h und die Vokalgemination, die auch in der erweiterten phonographischen Teilkompetenz des gutschrift-Tests Analyseeinheiten bilden. Dennoch schreibt es dort alle geforderten Wortbestandteile richtig.

Fallbeispiel 2 ist ein Schüler, der voraussichtlich auf das Gymnasium wechseln wird und die Bücherkategorie 101 bis 200 im Fragebogen angekreuzt hat. Im Rahmen der Analysen mit Hintergrundmerkmalen (vgl. Abschnitt 4.3) konnte gezeigt werden, dass Kinder mit angestrebter gymnasialer Laufbahn sowie bildungsnahem Elternhaus überdurchschnittlich gute Testergebnisse erzielen sowie Jungen im Peripheriebereich im Schnitt schlechter abschneiden als Mädchen. Das Kompetenzprofil des Kindes fügt sich damit gut in die Ergebnisse der Zusammenhangsanalysen ein.

### Fallbeispiel 3

An Fallbespiel 3 wird der Nutzen deutlich, die Schreibkompetenz eines Kindes nicht ausschließlich anhand der Anzahl richtig geschriebener Wörter zu beurteilen. In gutschrift wurden ausschließlich zwei Wörter und im SRT dreizehn Wörter korrekt verschriftlicht. In den Teilkompetenzen zeigen sich dagegen, neben den Schwächen, auch vier Bereiche, bei denen das Kind um die 70 Prozentpunkte erreicht. Dies sind die elementaren Kompetenzen von gutschrift und im SRT das phonographisch-silbische Prinzip sowie das der Wortbildung. Diese Ergebnisse lassen vermuten, dass Fallbespiel 3 grundlegende Einsichten in die Schriftsprache erworben hat, die noch ausgebaut und gefestigt werden sollten. Deutliche Förderbedarfe zeichnen sich bei den folgenden Teilkompetenzen ab: Der kleinste Wert, auch bezogen auf die Gesamtstichprobe, findet sich bei der erweiterten phonographischen Kompetenz mit nur einer richtig geschriebenen Analyseeinheit. Rund ein Drittel der Schreibungen sind bei der erweiterten grammatischen Kompetenz (36 Prozent) und dem Peripheriebereich (30 Prozent) sowie rund die Hälfte im morphologischen und wortübergreifenden Prinzip (jeweils 48 Prozent) richtig.

Auch dieses Kind hat den Kurzfragebogen ausgefüllt, aus dessen Auswertung entnommen werden kann, dass es sich um ein Mädchen mit Wechsel in die Realschule und einer im Haushalt verfügbaren Menge von 101-200 Büchern handelt.

### Fallbeispiel 4

Fallbeispiel 4 hat in allen Teilkompetenzen unterdurchschnittliche Werte, die aber zwischen den beiden Orthografietests stark schwanken. Dies deuten u. a. bereits die Prozentwerte auf Ganzwortebene an, die im gutschrift-Test im Vergleich zum SRT fast doppelt so groß ausfallen. Während in den SRT-Teilkompetenzen zweimal rund 15 und ansonsten zwischen 31 bis 35 Prozentpunkte gemessen werden konnten, liegen alle Leistungen in den Teilkompetenzen von gutschrift, wenn zum Teil auch nur knapp, über diesen Werten. Das Kind schreibt 36 Prozent der Indikatoren der beiden erweiterten Teilkompetenzen und 53 sowie 74 Prozent der beiden elementaren Kompetenzen richtig. Insgesamt betrachtet bereitet dem Viertklässler der gutschrift-Test weniger Schwierigkeiten als der SRT.

Das beschriebene Kompetenzprofil stammt von einem Jungen, der wahrscheinlich nach der Primarstufe auf die Hauptschule übergehen wird und bei der Frage nach der Bücheranzahl die Mittelkategorie (26 bis 100 Büchern) angeben hat.

### Fallbeispiel 5

Bei Fallbeispiel 5 handelt es sich um einen Grundschüler, dessen Prozentwerte in acht Teilkompetenzen zwischen 70 bis 93 liegen. Ausreißer bildet das wortübergreifende Prinzip mit 31 Prozent Richtigschreibungen. Dieser Wert fällt im Vergleich zur Gesamtstichprobe (84 Prozent) deutlich schwächer aus. Von den 29 Wörtern werden 20 Wörter fälschlicherweise klein geschrieben. Das Kind sollte daher in dem Bereich der Groß- und Kleinschreibung gefördert werden. Die elementare und die erweiterte grammatische Teilkompetenz des gutschrift-Tests umfassen ebenfalls Analyseeinheiten im Bereich der Groß- und Kleinschreibung. Hier erreicht das Kind 93 und 77 Prozent und liegt damit jeweils 6 Prozentpunkte oberhalb des Durchschnitts der Gesamtstichprobe. Da in gutschrift-diagnose die diktierten Wörter in dafür vorgesehene Lücken eingetragen werden mussten, ist das überdurchschnittliche Abschneiden des Kindes in den grammatischen Kompetenzen nicht generell widersprüchlich, sondern könnte über das Testformat erklärt werden. Zudem ist zu berücksichtigen, dass die beiden grammatischen Teilkompetenzen noch eine Reihe weiterer Items umfassen, die das Ergebnis beeinflussen.

Bei dem Kind handelt es sich um einen Jungen, der als weiterführende Schulform wahrscheinlich die Hauptschule besuchen wird und der in Bezug auf die Bücheranzahl die Antwortmöglichkeit 11 bis 25 Bücher angekreuzt hat. Damit weist er typische Determinanten für Bildungsbenachteiligung auf (vgl. Abschnitt 4.3).

### Resümee

Anhand der einzelnen Fallbeispiele konnte gezeigt werden, dass es Schülerinnen und Schüler mit erkennbaren Leistungsvariationen in den unterschiedlichen Teilkompetenzen des gutschrift-Tests und des SRT gibt. So wurden Schülerinnen und Schüler beschrieben, deren Kompetenzprofil sich dadurch auszeichnet, dass sie in einzelnen Teilkompetenzen vergleichsweise gut oder schlecht in Relation zu den weiteren Teilkompetenzen abschneiden (Fallbeispiele 1, 2 und 5). Anhand von Fallbeispiel 3 konnte ein Vorteil der differenzierten Auswertung auf Basis der Teilkompetenzen identifiziert werden. Ferner wurde mit Fallbeispiel 4 ein Schüler identifiziert, der in einem Rechtschreibtest bessere Ergebnisse erzielt als in dem anderen.

Diese hier beschriebenen oder mögliche ähnliche intraindividuelle Leistungsdifferenzen der Kinder – testabhängige sowie gute bzw. schlechte Leistungen in den Teilkompetenzen – könnten charakteristisch für eine ganze Gruppe von Schriftlernenden sein. Eine probabilistische Clusteranalyse soll daher im Folgenden angewendet werden, um die Frage systematischer Gruppengemeinsamkeiten zu beantworten.

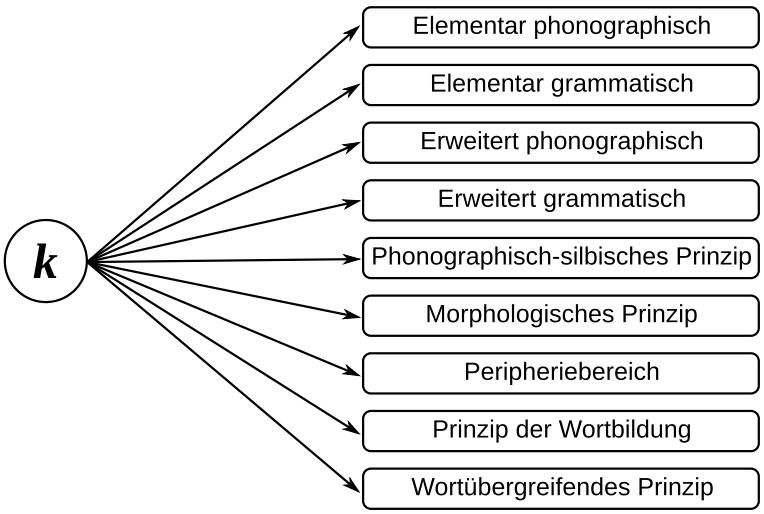



Abbildung 4.10: Modell der LPA mit den Teilkompetenzen als Indikatoren für die latente kategoriale Variable *k*

### 4.4.2 Latente Profile orthografischer Kompetenz

Bei dem in diesem Abschnitt verwendeten LPA-Messmodell wird von einer latenten kategorialen Variable *k* ausgegangen, die durch kontinuierliche manifeste Variablen gemessen wird. Abbildung 4.10 zeigt das entsprechende Pfaddiagramm. Als Indikatorvariablen (rechts dargestellt) werden die Leistungen in den Teilkompetenzen der beiden Rechtschreibtests genutzt. Die Pfeile von *k* zu den Teilkompetenzen bedeuten, dass das Antwortverhalten bzw. die Leistungswerte in den Klassen durch die latente Variable *k* erklärt werden: „It is particularly noteworthy that the causal flow is *from* the latent variable *to* the indicator variable, not the other way around. That is, observed indicator variables measure latent variables, but the observed indicator variables do not cause the latent variables.“ (Collins & Lanza, 2010, S. 4) Die Modellannahme der lokalen stochastischen Unabhängigkeit impliziert, dass die Zusammenhänge zwischen den Indikatorvariablen vollständig von der latenten Klassenvariable erklärt werden (J. Wang & X. Wang, 2012, S. 297). Als Kompetenzwerte wurden die WLE aus dem gemeinsamen Skalierungslauf der beiden Rechtschreibtests genutzt, bei dem die Größe der Stichprobe 547 Schülerinnen und Schüler beträgt (vgl. Abschnitt 4.2.3).

Mit der LPA wird eine weitere Möglichkeit genutzt, die Zusammenhänge zwischen den Teilkompetenzen der Tests zu analysieren. Ziel ist das Auffinden von Kindern mit vergleichbaren Kompetenzwerten, die in homogene Klassen gruppiert werden, sodass voneinander unterscheidbare Klassen mit unterschiedlichen Kompetenzprofilen entstehen. Es soll also geprüft werden, ob Profile identifiziert werden können, die, z. B. ähnlich den Einzelfallbeschreibungen der Kinder in Abschnitt 4.4.1, unterschiedliche Ausprägungen in den Teilkompetenzen aufweisen.

| Klassen | Para-meter | AIC | BIC | CAIC | ssa BIC | Entropy | *p* VLMRT | *p* LMRAT | *p* BLRT | Größe < 1% |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 18 | 59.341,27 | 59.418,75 | 59.436,75 | 59.361,61 | na | na | na | na | 0 |
| 2 | 28 | 57.253,43 | 57.373,96 | 57.401,96 | 57.285,08 | 0,894 | 0,000 | 0,000 | 0,000 | 0 |
| 3 | 38 | 56.516,85 | 56.680,42 | 56.718,42 | 56.559,79 | 0,895 | 0,001 | 0,001 | 0,000 | 0 |
| 4 | 48 | 56.172,94 | 56.379,55 | 56.427,55 | 56.227,18 | 0,896 | 0,248 | 0,251 | 0,000 | 0 |
| 5 | 58 | 56.026,05 | 56.275,71 | 56.333,71 | 56.091,59 | 0,862 | 0,180 | 0,183 | 0,000 | 0 |
| 6 | 68 | 55.970,68 | 56.263,38 | 56.331,38 | 56.047,52 | 0,856 | 0,179 | 0,181 | 0,000 | 0 |
| 7 | 78 | 55.927,14 | 56.262,89 | 56.340,89 | 56.015,29 | 0,871 | 0,580 | 0,583 | 0,000 | 1 |
| 8 | 88 | 55.885,80 | 56.264,59 | 56.352,59 | 55.985,25 | 0,859 | 0,215 | 0,217 | 0,000 | 1 |
| 9 | 98 | 55.843,53 | 56.265,37 | 56.363,37 | 55.954,28 | 0,843 | 0,135 | 0,140 | 0,000 | 1 |
| 10 | 108 | 55.806,96 | 56.271,84 | 56.379,84 | 55.929,01 | 0,855 | 0,191 | 0,196 | 0,000 | 1 |

ssa BIC = sample-size adjusted BIC
*p* VLRMT = *p* -Wert des Vuong-Lo-Mendell-Rubin Likelihood-Ratio-Tests
*p* LMRAT = *p* -Wert des Lo-Mendell-Rubin Adjusted-Likelihood-Ratio-Tests
*p* BLRT = *p* -Wert des Bootstrap-Likelihood-Ratio-Tests
Größe < 1% = Anzahl extrahierter Profile, die weniger als 1% der Stichprobe umfassen
na = not applicable

Tabelle 4.21: Vergleich der Modellanpassungen der Klassenlösungen

### Vergleich und Modellanpassung der Klassenlösungen

Mit den Daten des neundimensionalen Skalierungsmodells der beiden Rechtschreibtests wurden LPA-Modelle mit bis zu 10 Klassen explorativ berechnet. Alle hier dargestellten Analysen wurden mit der Software Mplus[17] (Version 7.11) durchgeführt. Die unterschiedlichen Klassenlösungen können anhand informationstheoretischer Maße und Likelihood-basierter Tests miteinander verglichen und bewertet werden (vgl. Abschnitt 3.3.3). In Tabelle 4.21 sind Informationen zu den berechneten Klassenmodellen und den Modellprüfgrößen zusammengestellt.

Der Entropy-Wert ist für alle Klassenlösungen zufriedenstellend. Er kann nicht direkt als ein Kriterium genutzt werden, um die Anzahl der Klassen zu bestimmen, gibt aber Hinweise auf die Güte der Klassifikation. Die Werte liegen alle über 0,8 und zeigen damit, dass die Klassen gut voneinander unterschieden werden können.

Aus den informationstheoretischen Maßen und den Likelihood-basierten Tests lassen sich unterschiedliche Implikationen ableiten. Der VLMRT und LMRAT sprechen für eine 3-Klassenlösung. Die *p*-Werte sind signifikant und betragen jeweils 0,001. Das geschätzte Modell mit 3 Profilen bildet damit die empirischen Daten statistisch bedeutsam besser ab als das Vergleichsmodell mit 2 Profilen. Die *p*-Werte von 0,248 bzw. 0,251 bei der 4-Klassenlösung zeigen hingegen, dass diese die Daten nicht besser abbildet als das 3-Klassenmodell. Die Kompetenzprofile der 3-Klassenlösung sind in Abbildung 4.11 dargestellt. Auf der Abszisse sind die Indikatorvariablen – die 9 Teilkompetenzen der

[17] http://www.statmodel.com

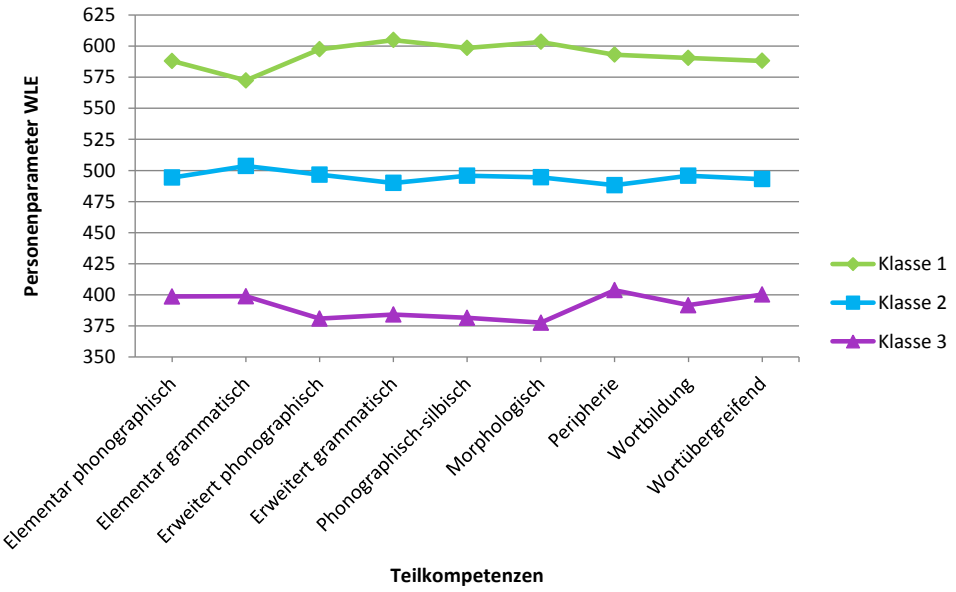



Abbildung 4.11: Kompetenzprofile der 3-Klassenlösung

beiden Rechtschreibtests – abgetragen und auf der Ordinate die mittleren Leistungswerte. Betrachtet werden sollten hier nicht die Verläufe der einzelnen Profile, da die Anordnung der Variablen auf der Abszisse „willkürlich“ ist (Rost, 2004, S. 149). Die Reihenfolge der Teilkompetenzen hätte auch anders gewählt werden können (z. B. alphabetisch oder erst die Teilkompetenzen des SRT und im Anschluss die des gutschrift-Tests). Stattdessen werden die Profilverläufe zwischen den einzelnen Klassen im Vergleich analysiert. Es können keine Überschneidungen, sondern ausschließlich Niveauunterschiede festgestellt werden. Die Klassen sind geordnet: In allen Teilkompetenzen einer Klasse werden im Vergleich zu den anderen Klassen entweder höhere oder niedrigere Kompetenzwerte erreicht. Schülerinnen und Schüler in Klasse 1 haben einen Leistungswert von rund 600, in Klasse 2 von rund 500 und in Klasse 3 von rund 400 Punkten. Es zeigt sich ein relativ paralleles Muster der Profile. Die LPA gruppiert die Schülerinnen und Schüler hierarchisch nach ihren Kompetenzwerten.

Beobachtet man die weiteren Verläufe der $p$-Werte des VLMRT und des LMRAT in Tabelle 4.21, nehmen sie bei der benachbarten 4-Klassenlösung zu und steigen im Anschluss mehrmals wieder ab und an. In der Simulationsstudie von Nylund et al. (2007, S. 563) zeigte sich, dass der BLRT, sobald der $p$-Wert einmal nicht signifikant wurde, dies auch konstant blieb. Wie Tabelle 4.21 zeigt, zeichnet dieser Wert für jede Modelllösung die benachbarte Lösung mit einer Klasse weniger als besser aus und ist damit ergebnislos.

Die Informationsindizes AIC, CAIC und saa BIC nehmen mit zunehmender Klassenanzahl ab und können damit ebenfalls nicht für die Wahl eines Klassenmodells herangezogen werden (vgl. Tabelle 4.21). Der BIC gibt hingegen Hinweise auf die Modellanpassung. Da sich der BIC in der Simulationsstudie als zuverlässiger und zu empfehlender Indikator für die Bestimmung der Klassenanzahl erwies (Nylund et al., 2007, S. 559) und dieser in seiner Anwendung bewährt und verbreitet ist (Magidson & Vermunt, 2004, S. 176), werden nun

| Klassen | geschätzte Klassengrößen | | manifeste Klassengrößen | |
|---|---|---|---|---|
| 1 | 39,12 | 7% | 40 | 7% |
| 2 | 136,03 | 25% | 131 | 24% |
| 3 | 159,93 | 29% | 168 | 31% |
| 4 | 125,14 | 23% | 125 | 23% |
| 5 | 68,83 | 13% | 66 | 12% |
| 6 | 17,95 | 3% | 17 | 3% |

Tabelle 4.22: Klassengrößen der 6-Klassenlösung

die Klassenlösungen mit niedrigem BIC-Wert detaillierter betrachtet. Die Maßzahl nimmt ab acht Klassen zu und hat ein Minimum von 56.262,89 bei der 7-Klassenlösung. Der Wert unterscheidet sich wenig von den benachbarten Lösungen, wie z. B. dem 6-Klassenmodell mit einem BIC von 56.263,38. Der Anteil der Schülerinnen und Schüler fällt aber bei dem 7-Klassenmodell nicht in allen Klassen substanziell aus. Eine Klasse beinhaltet nur vier Kinder, was 0,07 Prozent der Gesamtstichprobe entspricht. Die Klassengröße hat damit keine theoretische Bedeutsamkeit und ist nicht sinnvoll interpretierbar (J. Wang & X. Wang, 2012, S. 295). In solch einem Fall wird empfohlen, das Modell mit einer Klasse weniger auszuwählen (Samuelson & Raczynski, 2013, S. 310). Daher wird das sparsamere Modell mit 6 Klassen präferiert und im Folgenden näher analysiert.

Tabelle 4.22 enthält Informationen über die Anzahl der Kinder in den Profilen der 6-Klassenlösung. Der Anteil in den Klassen 2, 3 und 4 fällt am höchsten aus und beträgt zusammen etwa dreiviertel der Stichprobe. In der Tabelle sind zunächst die aufgrund des Modells geschätzten Klassengrößenparameter aufgelistet. Demnach gehören der ersten Klasse 7, der zweiten Klasse 25, der dritten Klasse 29, der vierten Klasse 23, der fünften Klasse 13 und der sechsten Klasse 3 Prozent an. Die Klassengrößen auf Grundlage der manifesten Zuordnung weichen leicht von den geschätzten ab. Nach Geiser (2011, S. 249) sind Abweichungen üblich, da bei der manifesten Zuordnung von Personen auf Basis der maximalen Klassenzuordnungswahrscheinlichkeiten Schätzfehler auftreten.

Neben dem Entropy-Wert und den Klassengrößen können die mittleren Klassenzuordnungswahrscheinlichkeiten als ein weiteres Maß für die Zuverlässigkeit der Klassifikation der Personen herangezogen werden. Sie sind in Tabelle 4.23 wiedergegeben. Die Werte auf der Hauptdiagonale liegen bei allen Profilen über 0,8 und bei vier Profilen über 0,9 und fallen damit sehr gut aus. Die Schülerinnen und Schüler werden mit einer „Treff“-Sicherheit um 90 Prozent den Klassen zugeordnet. Über alle Klassen hinweg betrachtet fallen die „Unschärfen“ gering aus. Sie betreffen ausschließlich die direkt benachbarten Klassen und liegen jeweils unter 10 Prozent. Beispielsweise beträgt die Wahrscheinlichkeit von Kindern in der ersten Klasse, auch der zweiten Klasse anzugehören, 7 Prozent.

| Klassen und -zugehörigkeit | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | **0,93** | 0,07 | 0,00 | 0,00 | 0,00 | 0,00 |
| 2 | 0,01 | **0,93** | 0,06 | 0,00 | 0,00 | 0,00 |
| 3 | 0,00 | 0,07 | **0,86** | 0,07 | 0,00 | 0,00 |
| 4 | 0,00 | 0,00 | 0,06 | **0,88** | 0,06 | 0,00 |
| 5 | 0,00 | 0,00 | 0,00 | 0,06 | **0,91** | 0,04 |
| 6 | 0,00 | 0,00 | 0,00 | 0,00 | 0,08 | **0,92** |

Tabelle 4.23: Mittlere Klassenzuordnungswahrscheinlichkeiten der 6-Klassenlösung

### Profilbeschreibung

Die Profilverläufe der 6-Klassenlösung sind in Abbildung 4.12 dargestellt. Auch hier gilt gleiches wie zuvor bereits bei der 3-Klassenlösung beobachtet werden konnte: Die Profile sind geordnet und die LPA hat homogene Gruppen der Kinder auf Basis ihrer Kompetenzwerte gebildet. Sie betragen durchschnittlich rund:

- 650 Punkte in Klasse 1
- 570 Punkte in Klasse 2
- 510 Punkte in Klasse 3
- 450 Punkte in Klasse 4
- 385 Punkte in Klasse 5
- 320 Punkte in Klasse 6

Die Schülerinnen und Schüler in Klasse 1 stellen die leistungsstärkste Gruppe dar und befinden sich damit auf einem herausragenden Kompetenzniveau. Zur leistungsschwächsten Klasse weisen sie einen durchschnittlichen Vorsprung von 330 Punkten auf. Die Kinder in Klasse 2 erreichen ebenfalls überdurchschnittliche Ergebnisse, während die Kinder in den Klassen 4 bis 6 unterdurchschnittliche Kompetenzwerte aufweisen. Die Gruppe der Schülerinnen und Schüler in Klasse 3 liegt mit 510 Punkten nahe dem Skalenmittelwert.

Die Profilverläufe der 6-Klassenlösung sind also auch nach Leistung geordnet. Gleichzeitig sind sie aber variationsreicher als die der 3-Klassenlösung. In Klasse 1 existieren innerhalb der grammatischen Teilkompetenzen des gutschrift-Tests mit 80 Punkten die deutlichsten relativen Unterschiede in den Kompetenzwerten. Die Schülerinnen und Schüler dieser Gruppe erzielen in der elementaren Kompetenz 611 und in der erweiterten Kompetenz 691 Punkte. Weitere größere Differenzen ergeben sich in den beiden leistungsschwächsten Klassen 5 und 6 mit 45 und 49 Punkten. In Klasse 5 sind relative Schwächen im morphologischen Prinzip (367 Punkte) und Stärken in der elementaren grammatischen Kompetenz

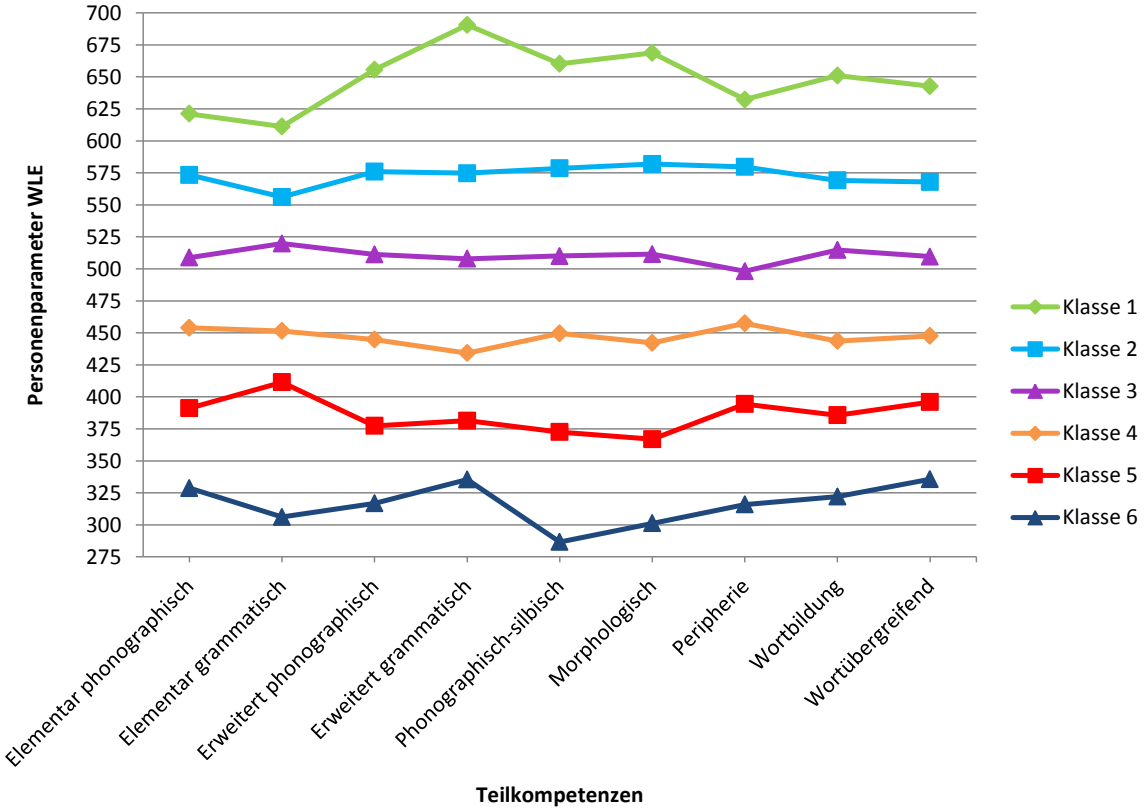



Abbildung 4.12: Kompetenzprofile der 6-Klassenlösung

(412 Punkte) erkennbar. In Klasse 6 sind diese zwischen dem phonographisch-silbischen Prinzip (287 Punkte) und der erweiterten grammatischen Teilkompetenz bzw. dem wortübergreifenden Prinzip (jeweils 335 Punkte) vorhanden. Die Profilverläufe der weiteren Klassen 2 bis 4 sind ausgeglichener. Es handelt sich hierbei um Klassen, die über Schülerinnen und Schüler mit eher mittleren Kompetenzwerten gebildet worden sind und deren Leistungen in den neun orthografischen Teilkompetenzen ein homogeneres Bild ergeben. Innerhalb dieser Profile existieren relative Differenzen zwischen den Leistungswerten in den Teilkompetenzen von durchschnittlich 24 Leistungspunkten.

Es zeigt sich, dass insbesondere bei den Randklassen stärkere Leistungssprünge in den Teilkompetenzen erkennbar sind. Dieser Befund ist analog zu den im Rahmen von IGLU/TIMSS getätigten LPA (vgl. Abschnitt 3.3.3). Dort weisen ebenfalls die Profilverläufe der leistungsstärksten und -schwächsten Typen die größten Leistungsunterschiede in den beobachteten Subdomänen der Kompetenzbereiche Lesen, Mathematik und Naturwissenschaften auf (Bos et al., 2012a, S. 253). Sie betragen maximal rund 40 Punkte innerhalb eines Typs bei der fachspezifischen Subskala Mathematik. Die hier für den Bereich der Orthografie ermittelten Differenzen in den Profilverläufen fallen teilweise größer aus. Insbesondere sind hier die Unterschiede in den Profilverläufen der Klassen 1, 5 und 6 zu nennen (vgl. Abbildung 4.12). So sind diese für die leistungsstärkste Klasse 1 (80 Punkte) doppelt so groß im Vergleich zur maximal ermittelten Differenz in IGLU/TIMSS

(ca. 40 Punkte). Auch die Differenzen in den beiden leistungsschwächsten Klassen (45 und 49 Punkte) übersteigen diesen Wert leicht.

#### Profilinterpretation

Da die beiden Randprofile am variationsreichsten sind, sollen diese im Folgenden betrachtet werden. Auffällig erscheint im Profil der Klasse 1 das Abschneiden in den elementaren (621 und 611 Punkte) im Vergleich zu den erweiterten (656 und 691 Punkte) gutschrift-Teilkompetenzen. Dies zeigt, dass sie, im Vergleich zu den restlichen Kindern in der Stichprobe, die der Theorie nach ausgewiesenen komplexen Rechtschreibphänomene, wie z. B. Kürze- und Dehnungszeichen zur Markierung der kurzen bzw. langen Vokale (z. B.: Dehnungs-h und -e) oder die Großschreibung von abstrakten Nomen und Nominalisierung noch deutlicher besser bewältigen als die elementaren Teilkompetenzen. Dies könnte zu der Interpretation führen, dass sich die besten Schülerinnen und Schüler auf solch einem hohen schriftsprachlichen Niveau bewegen, dass sie bereits die Mehrheit an schwierigen orthografischen Phänomenen erfolgreich in ihr Schriftprodukt integrieren, während der Fokus der Kinder in den anderen Typen noch auf den einfacheren Phänomenen zu liegen scheint.

Im Profilverlauf der Klasse 6 fallen relative Stärken in der erweiterten grammatischen Teilkompetenz (335 Punkte) und dem wortübergreifenden Prinzip (336 Punkte) auf. Obwohl die erweiterte grammatische Teilkompetenz hoch mit dem morphologischen Prinzip korreliert und ähnliche Analyseeinheiten umfasst (vgl. Abschnitte 2.4 und 4.2.3), liegen diese beiden Teilkompetenzen bei dieser Gruppe nicht auf einem Niveau. Stattdessen weisen die Schülerinnen und Schüler im morphologischen Prinzip (301 Punkte) und im phonographisch-silbischen (287 Punkte) relativ zu den anderen Schülerinnen und Schülern die größten Schwierigkeiten auf. In den beiden Kernbereichen des SRT wie auch in der elementaren grammatischen Teilkompetenz (306 Punkte) von gutschrift erreichen die Schülerinnen und Schüler damit relativ niedrigere Kompetenzwerte als in den erweiterten Teilkompetenzen (317 und 335 Punkte) und im Peripheriebereich (316 Punkte). Hier könnte geschlussfolgert werden, dass es sich um eine Gruppe von Schülerinnen und Schülern handelt, die grundlegende Systematiken und Regeln des Schriftsystems nicht verstehen und Wortschreibungen eher durch z. B. Auswendiglernen (re-)produzieren. Bei der Interpretation ist insbesondere bei dieser Klasse zu bedenken, dass sich aufgrund der kleinen Schüleranzahl im Profilverlauf höhere Variationen einstellen können.

### 4.4.3 Zusammenschau

Im Rahmen der Einzelfallanalysen (vgl. Abschnitt 4.4.1) konnten auf Individualebene Leistungsvariationen in den unterschiedlichen Teilkompetenzen des gutschrift-Tests und des SRT erkannt werden. In der LPA zeigten sich keine Gruppen von Kindern, die sich aufgrund von variierenden Leistungen in den Teilkompetenzen in voneinander unterscheidbare Schülergruppen klassifizieren lassen. Systematische Klassenunterschiede konnten

allein auf Basis der Rangordnung der Kinder in den Kompetenzwerten analysiert werden. Im Hinblick auf die ermittelten und zum Teil hohen latenten Zusammenhänge zwischen den Teilkompetenzen (vgl. Tabelle 4.14) ist dieser Befund nicht erwartungswidrig. Die Kompetenzprofile unterscheiden sich durch die Höhe des erzielten Werts in den Teilkompetenzen voneinander. Die Profile überschneiden sich selber nie. Dies zeigten sowohl die Profilverläufe der 3-Klassenlösung als auch die der 6-Klassenlösung.

Über die Modellgeltungstests konnte keine eindeutige Aussage für eine der beiden Klassenlösungen getroffen werden. Die 6-Klassenlösung wurde auf Basis eines etablierten Modellwahlkriteriums und in Kombination mit inhaltlichen Überlegungen verstärkt betrachtet. Bei dieser Klassenlösung weisen die Kinder in den Klassen, deren Personenfähigkeiten um den Skalenmittelwert liegen, relativ ausgewogene Profilverläufe auf. Dahingegen zeigen sich bei den beiden Klassen an den zwei Extrema der Leistungsskala die stärksten Variationen in den Kompetenzwerten. Diese Beobachtungen aus ZuRecht spiegeln die Befunde aus IGLU/TIMSS wider. In der Studie sind die Klassen ebenfalls auf Basis der Leistungswerte geordnet. Die hier vorgefundenen mittleren klasseninternen Differenzen übersteigen die in IGLU/TIMSS dokumentierten Unterschiede in den Subdomänen. So sind beispielsweise in der leistungsstärksten Klasse relative Unterschiede in den Kompetenzwerten auszumachen, die doppelt so groß sind wie die maximal ermittelte Differenz in IGLU/TIMSS.

# 5 Fazit und Ausblick

In diesem Kapitel werden die gestellten Forschungsfragen aus Abschnitt 1.2 anhand der in Kapitel 4 erfolgten Datenanalysen sowie der theoretischen Gegenüberstellung aus Abschnitt 2.4 zusammenfassend beantwortet (Abschnitt 5.1). Dabei werden die wichtigsten Ergebnisse aufgegriffen und auf Basis der Datenskalierungen Hinweise auf mögliche zukünftige Weiterentwicklungen der beiden Rechtschreibtests gegeben. Vervollständigt wird die Arbeit durch die Diskussion der im Rahmen von ZuRecht gewonnenen Erkenntnisse und darauf aufbauender weiterer Forschungsfragen (Abschnitt 5.2).

## 5.1 Beantwortung der Forschungsfragen und Implikationen

### Forschungsfrage I

Die erste Forschungsfrage thematisiert den orthografischen Kompetenzstand der Schülerinnen und Schüler aus ZuRecht und soll im Folgenden beantwortet werden. Dabei werden die Ergebnisse auf Ganzwort- und Teilkompetenzebene zusammenfassend geschildert und um Angaben zu Variantenschreibungen sowie Stärken und Schwächen bei Wortschreibungen ergänzt.

Die Berechnung univariater Statistiken auf Wortebene zeigte, dass die Schülerinnen und Schüler Schwierigkeiten mit der Verschriftung der Testwörter beider Diktate hatten. Die durchschnittliche Fehleranzahl und Leistungsstreuung fallen beim gutschrift-Test, bedingt durch die längeren und damit komplexeren Wörter (insbesondere Komposita), etwas höher aus. Es gibt kein Kind, das den gutschrift-Test oder den SRT vollständig fehlerfrei verschriftet hat. Gleichzeitig ist eine Vielzahl von Variantenschreibungen vorhanden (vgl. Tabellen 4.1 auf Seite 112 und 4.2 auf Seite 113). Auf ein Diktatwort entfallen, je nach Test, durchschnittlich 65 (gutschrift) bzw. 45 (SRT selektiert) unterschiedliche Schreib-

varianten. Die Schreibungen der Testwörter sind so mannigfaltig, dass damit eine große Verunsicherung der Schülerinnen und Schüler zum Ausdruck kommt.

Besondere Verunsicherungen waren mit den abgeleiteten Testwörtern empfindlich (gutschrift) und Schnurrbarthaaren (SRT) verbunden (vgl. Tabelle 4.3 auf Seite 114). Bei dem ersten Testwort sind Defizite im lautanalytischen Erschließen von ⟨m⟩, der Verschriftung der Affrikate /pf/ sowie der Auslautverhärtung des ⟨d⟩ erkennbar, während das Suffix ⟨lich⟩ wenige Probleme verursacht. Bei dem zweitgenannten Diktatwort bereiten vor allem die Konsonanten- und Vokalgraphemgemination Schwierigkeiten und es findet mehrfach eine falsche Schreibung des Phonems /t/ durch ⟨d⟩ anstatt ⟨t⟩ statt. Sicher beherrscht werden hingegen die Schreibung des Wortanfangs /ʃ/ als ⟨sch⟩ sowie das Suffix ⟨en⟩. Schnurrbarthaaren bildet gleichzeitig das schwierigste Wort des SRT; bei gutschrift lautet es Schnellste. Hier scheitern die Kinder hauptsächlich an der Substantivierung. Durch das Fließtextdiktat des SRT zeigte sich, dass kleine Wörter wie Artikel, Konjunktionen, Präpositionen und Pronomen fast durchgängig richtig geschrieben werden. Darunter fällt auch das Wort du mit den meisten Richtigschreibungen. Im gutschrift-Test ist das leichteste Wort viele.

Über die Aufteilung der Stichprobe in das 15. und 85. Perzentil konnte ermittelt werden, dass die leistungsschwächsten 15 Prozent der Viertklässler maximal 34 Prozent (gutschrift) bzw. 42 Prozent (SRT selektiert) der Testwörter richtig schreiben, während die leistungsstärksten 15 Prozent den Großteil der Testwörter (mindestens etwa 80 Prozent) normgerecht verschriften (vgl. Tabelle 4.2 auf Seite 113). Damit zeigt sich eine deutliche Heterogenität der Rechtschreibkompetenz innerhalb der vierten Jahrgangsstufe.

Die Häufigkeitsauszählung der Teilkompetenzen von gutschrift und dem SRT zeigen differenzielle Ergebnisse (vgl. Tabelle 4.13 auf Seite 135). Die drei schwierigsten Subskalen stellen die erweiterte phonographische und die erweiterte grammatische Kompetenz des gutschrift-Tests (Prozent-korrekt-Werte von 77 bzw. 71) sowie der Peripheriebereich des SRT (74 Prozent) dar. Der Mittelwert an Richtigschreibungen fällt hingegen u. a. für die elementaren Teilkompetenzen des gutschrift-Tests (88 und 87 Prozent) sowie das phonographisch-silbische Prinzip des SRT (91 Prozent) hoch aus. Hier werden also fast alle Struktureinheiten richtig geschrieben. Die Befunde zeigen, dass der Großteil der Schülerinnen und Schüler über grundlegende orthografische Fähigkeiten verfügt. Gleichzeitig können die theoretisch von den Kompetenzmodellen angenommenen Schwierigkeitsgrade bestätigt werden. Die Indikatoren der elementaren Teilkompetenzen und die der beiden Kernbereiche werden häufiger richtig geschrieben als die der erweiterten Teilkompetenzen und die des Peripheriebereichs.

### Forschungsfrage II

Im Mittelpunkt der zweiten Forschungsfrage steht die Validierung der linguistischen Rechtschreibkompetenzmodelle von gutschrift und dem SRT. Dafür wurden die Daten mit einparametrigen logistischen IRT-Modellen skaliert, um die theoretischen Vorannahmen der Testkonzeptionen zu prüfen. Die Frage nach der Qualität der Items und Güte der Gesamtmodelle wird im Folgenden beantwortet.

Bei der Skalierung des gutschrift-Tests konnte zunächst gezeigt werden, dass das einparametrige dichotome Raschmodell auf die Datenstruktur angewendet werden kann (vgl. Tabelle 4.4 auf Seite 118). Auf diesem Modell basieren daher für gutschrift alle weiteren Berechnungen. Im Rahmen des Itemreviews mussten über 35 Prozent der Struktureinheiten im gutschrift-Test und knapp 20 Prozent aus dem SRT ausgeschlossen werden, da sie sich als nicht verträglich mit dem Modell erwiesen. Sie wiesen ungenügende Eigenschaften in Form eines unzureichenden Itemfits und/oder einer geringen Trennschärfe auf. In der elementaren grammatischen Teilkompetenz von gutschrift verblieben infolgedessen mit 15 Items nur wenige Analyseeinheiten. Diese Teilkompetenz sollte zukünftig mit neuen angemessenen Items aufgefüllt werden, um eine konsistente und präzise Schätzung der Personenparameter zu ermöglichen. Die Gegenüberstellung der Personenfähigkeiten (orthografische Kompetenz der Schülerinnen und Schüler) und Itemschwierigkeiten (Analyseeinheiten der Teilkompetenzen) ergab eine linksschiefe Verteilung, die über die Wright-Map visualisiert wurde. Die durchschnittlichen Lösungswahrscheinlichkeiten fallen höher als 50 Prozent aus, da die Personenparameter im Mittel die Aufgabenparameter übersteigen. Insgesamt messen beide Tests damit etwas zu leicht und differenzieren im unteren Kompetenzbereich stärker als im oberen (vgl. Abbildungen 4.2 auf Seite 120 und 4.5 auf Seite 129). Hier sollten die Rechtschreibtests zukünftig durch schwierigere Items angereichert werden, um das gesamte Leistungsspektrum zu erfassen.

Im Rahmen der Dimensionalitätsprüfung der theoriekonformen Rechtschreibkompetenzmodelle von gutschrift-diagnose (vier Teilkompetenzen) und dem SRT (fünf Teilkompetenzen) wurden die mehrdimensionalen Modelle konkurrierenden eindimensionalen Modellvarianten, sogenannten Generalfaktormodellen, gegenübergestellt. Der Modellvergleich erfolgte mittels informationstheoretischer Maße sowie Likelihoodquotiententest und Chi-Quadrat-verteilter Prüfstatistik, um die Passung zwischen theoretischer und empirischer Modellstruktur zu quantifizieren. Die Informationsindizes wiesen die mehrdimensionalen Modelle als relativ beste aus. Eine inferenzstatistisch abgesicherte Aussage konnte mittels Likelihoodquotient und Überführung in eine $\chi^2$-verteile Prüfgröße gewonnen werden. Das Ergebnis des Tests war signifikant: Die Modellprüfung ergab für die in den Daten enthaltenen Informationen eine höhere Erklärungskraft des vierdimensionalen Modells von gutschrift und des fünfdimensionalen Modells vom SRT (vgl. Tabellen 4.5 auf Seite 122 und 4.8 auf Seite 131). Demzufolge ergibt sich eine empirische Evidenz für die differenziellen Rechtschreibkompetenzmodelle.

Diese Ergebnisse auf multivariater Ebene decken sich teilweise mit den vorgefundenen korrelativen Strukturen. Sie zeigen für den gutschrift-Test hohe latente Zusammenhänge zwischen den Teilkompetenzen, die im Rahmen von 0,87 bis 0,96 liegen (vgl. Tabelle 4.6 auf Seite 124). Sie sind aber, verglichen mit den Interkorrelationen in z. B. PISA, die als Referenzwerte hinzugezogen worden sind, um die Enge des statistischen Zusammenhangs beurteilen zu können, nicht unüblich. Die Teilkompetenzen des SRT liegen zwischen 0,71 und 0,96 (vgl. Tabelle 4.9 auf Seite 131) und können im Vergleich zu gutschrift besser voneinander unterschieden werden bzw. sich als eigenständiger erweisen. Dies betrifft insbesondere die Zusammenhänge des Peripheriebereichs und des wortübergreifenden Prinzips. Die Varianzanteile betragen für diese beiden Teilkompetenzen 50 bis 69 Pro-

zent (Peripheriebereich) sowie 50 bis 74 Prozent (wortübergreifendes Prinzip). Hieraus kann geschlossen werden, dass Ausnahmeschreibungen und die Großschreibung andere Anforderungen stellen. Bezüglich der Korrelation zwischen den Teilkompetenzen des Kernbereichs (phonographisch-silbisches und morphologisches Prinzip) und ähnlich ausfallender Zusammenhänge dieser beiden Prinzipien mit den weiteren Teilkompetenzen des SRT sollte zukünftig dessen Struktur überdacht werden. Eine vierdimensionale Modellierung (vgl. Tabellen 4.10 und 4.11 auf Seite 132) wäre hierbei denkbar und könnte vertieft betrachtet werden. Ebenfalls bei gutschrift sollte, auf Basis der korrelationsstatistischen Analyse der Teilkompetenzen, die Struktur überarbeitet werden. Ein zweidimensionales fähigkeitsbasiertes Modell (1. Dimension: phonographische Teilkompetenz, 2. Dimension: grammatische Teilkompetenz), das zusätzlich berechnet wurde, erwies sich allerdings mit einem Zusammenhang von 0,98 als unzulänglich (vgl. Abschnitt 4.2.1).

Durch die arbiträre Itemklassifikation, bei der die Struktureinheiten beliebig den Teilkompetenzen zugeordnet worden sind, zeigte sich insbesondere für den SRT eindrucksvoll, dass die Korrelationen zwischen den Teilkompetenzen nicht zufällig zustande kommen, sondern das Ergebnis theoretischer Formulierungen darstellen (vgl. Tabelle 4.12 auf Seite 133).

Die im Rahmen dieser Arbeit erstmalig vorgenommene Validierung des gutschrift-Kompetenzmodells zeigt Befunde, die dafür sprechen, die Modellierung der orthografischen Kompetenz, wie sie in ihrer jetzigen Form besteht, zu verbessern. Der Bedarf erstreckt sich dabei auf das Anreichern der Teilkompetenzen um schwierigere Items und im speziellen auf die Erhöhung der Itemanzahl der elementaren grammatischen Teilkompetenz. Weitaus relevanter erscheint aber die Notwendigkeit des Senkens der hohen Zusammenhänge zwischen den Subskalen, damit diese als eigenständig gelten können. Die nach neuen Verfahrensweisen (unter Verwendung des SRT-Editors) wiederholte Validierung des SRT-Kompetenzmodells für die vierte Jahrgangsstufe spricht für eine Zusammenlegung der Kernbereiche. Inwieweit ein vierdimensionales Modell inhaltlich und didaktisch plausibel ist, sollte von Experten der Linguistik weiter diskutiert werden. Zudem sollten auch hier alle Teilkompetenzen um schwierigere Struktureinheiten ergänzt werden.

### Forschungsfrage III

Zur Beantwortung der dritten Forschungsfrage, die den Vergleich der beiden Tests thematisiert, werden der gutschrift-Test und der SRT nicht mehr einzeln betrachtet, sondern nach schrifttheoretischen und psychometrischen Kriterien gegenübergestellt. Hierbei werden die Zuordnungen der Analyseeinheiten, die Reliabilitäten der Teilkompetenzen sowie die testübergreifenden Korrelationen beleuchtet.

Mittels quantitativer Inhaltsanalyse wurden die rechtschriftlichen Analyseeinheiten, also die Indikatoren zur Operationalisierung der Teilkompetenzen, quantifiziert (vgl. Tabelle 2.11 auf Seite 61). Auf diese Weise war es möglich, größere Überschneidungen bei der elementaren phonographischen Kompetenz des gutschrift-Tests und dem phonographisch-silbischen Prinzip des SRT sowie der erweiterten grammatischen Kompetenz (gutschrift)

und dem morphologischen Prinzip (SRT) zu identifizieren. Die verschiedenen Funktionen, die die Rechtschreibphänomene im Schriftsystem laut Testkonzeptionen einnehmen sowie die Gründe für die Zuordnungen zu den Teilkompetenzen wurden erläutert, indem die Tests in ihren linguistischen Hauptannahmen und -merkmalen abgrenzend voneinander dargestellt und zum Teil kritisch beleuchtet wurden. Als zentral erwiesen sich dabei die folgenden Punkte:

- Das gutschrift-Konzept ist gekennzeichnet durch die Orientierung an curricularen Vorgaben. Die Auswahl der Indikatoren sowie die Zuordnung zu den Teilkompetenzen erfolgt nach den Inhalten dieser Lehrpläne. Die Unterteilung der Fähigkeiten in phonographisch und grammatisch basiert auf der Betrachtung der Abweichung von lauttreuen Schreibungen. Die Niveaus elementar und erweitert gründen auf Schwierigkeitsanforderungen und werden in Form von Stufenfolgen auf den Unterricht angewendet.
- Das SRT-Konzept ist charakterisiert durch die Graphematik nach Eisenberg und die didaktische Umsetzung nach Hinney. Hier werden ein Kern- und ein Peripheriebereich der Wortschreibungen unterschieden. Der Kernbereich ist durch das phonographisch-silbische Prinzip, das morphologische Prinzip sowie das Prinzip der Wortbildung herleitbar. In diesem Zusammenhang wird die Systematik der Orthografie sowie ihre leserfreundliche Funktion betont. Als zentrales Mittel zur Erklärung rechtschreiblicher Zusammenhänge wird die Rolle der Silbe herausgestellt.

Die gemeinsame neundimensionale Skalierung mit den Teilkompetenzen von gutschrift und dem SRT wurde durch das besondere Erhebungsdesign von ZuRecht ermöglicht. Es zeichnet sich durch den simultanen Einsatz der beiden Tests in einer identischen Stichprobe aus. Die korrelationsstatistischen Befunde verifizieren die inhaltsanalytische Quantifizierung der Analyseeinheiten (vgl. Tabelle 4.14 auf Seite 137). Hier zeigt sich theoriekonform der stärkste Zusammenhang (0,94) zwischen der erweiterten grammatischen Teilkompetenz und dem morphologischen Prinzip. Die gemeinsame aufgeklärte Varianz beträgt 88 Prozent. Diese beiden Teilkompetenzen erfassen damit äußerst ähnliche Informationen. Die niedrigsten Korrelationen ergaben sich bei dem Peripheriebereich (0,70 - 0,79) und dem wortübergreifenden Prinzip (0,71 - 0,83). Diese beiden Teilkompetenzen des SRT erwiesen sich bereits in dem fünfdimensionalen Modelllauf als selbstständigste Größen. Die neundimensionale Skalierung bestärkt damit zusätzlich diese Ausdifferenzierung.

Die Reliabilitäten konnten über die testübergreifende Skalierung in unmittelbare Beziehung zueinander gesetzt werden und ergaben für gutschrift und den SRT vergleichbar gute Werte, die durchgängig über 0,8 liegen (vgl. Tabelle 4.15 auf Seite 139). Die orthografischen Teilkompetenzen der Kinder werden mit den Analyseeinheiten beider Tests auf Basis dieser Statistik genau erfasst, womit kein Vorteil für einen Test ausgemacht werden kann.

### Forschungsfrage IV

Die Reproduktion der Ergebnisse des SRT und ihre Verortung in der IGLU-Hauptuntersuchung sind Bestandteil der vierten Forschungsfrage und werden durch den Einsatz der identischen Testversion in beiden Erhebungen ermöglicht. Hier soll durch den größeren Referenzrahmen zum einen der orthografische Kompetenzstand der Schülerinnen und Schüler aus ZuRecht eingeschätzt, und zum anderen geprüft werden, ob die in ZuRecht vorgefundene Kompetenzstruktur verifiziert werden kann. Daher wurden im Rahmen dieser Arbeit die SRT-Rechtschreibdaten der IGLU-HE erstmals ausgewertet und veröffentlicht.

Die Auszählung der Schreibungen der Analyseeinheiten indiziert, dass die Kinder aus der bundesweit repräsentativen Stichprobe marginal unterschiedlich in den Teilkompetenzen des SRT abschneiden (vgl. Tabelle 4.18 auf Seite 143). Die maximale Differenz beträgt 4 Prozentpunkte. Dementsprechend erweist sich beispielsweise der Peripheriebereich weiterhin als die für die Viertklässler schwierigste Teilkompetenz und das phongraphisch-silbische Prinzip als die leichteste. Es lässt sich ableiten, dass die in ZuRecht getesteten Schülerinnen und Schüler ein für Deutschland durchschnittliches Kompetenzniveau aufweisen. Daher ist es erwartungskonform, dass die Schwierigkeitsverteilung der Items analog zu ZuRecht linksschief ausfällt (vgl. Abbildung 4.6 auf Seite 141). Die mittleren Personenfähigkeiten übersteigen die mittleren Itemschwierigkeiten. Im Rahmen der Itemanalyse mussten 16 Prozent der Struktureinheiten eliminiert werden. Dieser Anteil fällt damit etwas kleiner als in ZuRecht aus.

Die formalstatistischen Modellgeltungstests auf Grundlage der Likelihood-Statistiken für den SRT sprechen ebenfalls dafür, dass sich die Rechtschreibdaten mit der theoretisch angenommenen komplexen Kompetenzstruktur besser beschreiben lassen als mit einem Generalfaktormodell. Der errechnete Wert der Prüfgröße liegt hier weit oberhalb der kritischen Grenze (vgl. Tabelle 4.16 auf Seite 142).

Die innerhalb von ZuRecht vorgefundenen Zusammenhänge zwischen den SRT-Subskalen können ebenfalls reproduziert werden. Es lassen sich vergleichbare Interkorrelationen beobachten (vgl. Tabelle 4.17 auf Seite 142). Sie differieren um maximal 0,06 Punkte. Auch hier erweisen sich der Peripheriebereich und das wortübergreifende Prinzip aus psychometrischer Sicht als am eigenständigsten. Der Varianzanteil liegt zwischen 53 und 76 Prozent. Ebenso korrelieren mit 0,96 die beiden Kernbereiche hoch miteinander, was dem ermittelten Wert aus ZuRecht identisch entspricht.

### Forschungsfrage V

Innerhalb der fünften Forschungsfrage wurden die Zusammenhänge der Variablen Geschlecht, weiterführende Schulform und Bildungshintergrund betrachtet. Diese Analysen beschränken sich nicht auf die Auswertung der Wortfehler, sondern beziehen die differenziellen Teilkompetenzen mit ein. Für die Analyse wurden die Parameterwerte der Personenfähigkeiten jedes Kindes in Form von Weighted-Likelihood-Estimates genutzt

und auf einen Mittelwert von 500 und eine Standardabweichung von 100 normalisiert. Die Tragweite der Mittelwertsunterschiede wurde durch die Berechnung von Effektstärken nach Cohen einschätzbar.

Bei den Genderanalysen schnitten die Mädchen auf Wort- und Teilkompetenzebene in beiden Tests besser als die Jungen ab (vgl. Abbildung 4.7 auf Seite 148). Die Mittelwertsunterschiede sind beim SRT stärker als bei gutschrift ausgeprägt. Daher wären weiterführende Analysen zu geschlechtsspezifischen Wörtern mittels differential item functioning aufschlussreich, um ggf. vorhandene Vor- und Nachteile der verwendeten Diktatinhalte für die beiden Gruppen aufzudecken. Die standardisierten Mittelwertsunterschiede erwiesen sich für den Peripheriebereich als auffällig hoch. Hier wurde ein mittlerer bis großer Effekt von $d = 0,67$ zu Gunsten der Mädchen ermittelt. Demnach fallen ihnen die Schreibungen von Ausnahme- und Merkwörtern deutlich einfacher als den Jungen. Mögliche Ursachen könnten in den besseren Leseleistungen oder dem erhöhten außerschulischen Leseverhalten der Schülerinnen liegen. Ebenfalls könnte die Beherrschung der restlichen Teilkompetenzen eine Voraussetzung für die kognitiven Anforderungen des Peripheriebereichs darstellen, und damit Einfluss auf die Fähigkeiten in diesem Bereich ausüben.

Die Zusammenhangsanalysen von Rechtschreibkompetenz und weiterführender Schulform zeigten in allen Subskalen ähnlich große Kompetenzmittelwerte innerhalb der jeweiligen Schulform. Damit konnten keine schulformcharakteristischen Teilkompetenzen identifiziert werden. Stattdessen konnte eine monotone Steigung der durchschnittlichen Kompetenzwerte von der Hauptschule zum Gymnasium hin beobachtet werden (vgl. Abbildung 4.8 auf Seite 151). Die Kompetenzvorsprünge der Gymnasiasten gegenüber den Hauptschülern fallen auf Gesamtwortebene für beide Tests mit einer Effektstärke von $d \approx 2$ am höchsten aus. Dieser Wert demonstriert einen bedeutenden Kompetenzvorsprung bzw. -rückstand der Schülerinnen und Schüler in Abhängigkeit von der erwarteten weiterführenden Schulform. Die Wahrscheinlichkeit eines höheren Schulformbesuchs steht damit in einem empirisch nachgewiesenen Zusammenhang zu der Höhe der orthografischen Kompetenz. Von der Herleitung einer kausalen Abhängigkeit ist aber zwingend abzusehen. Stattdessen wären weiterführende Studien empfehlenwert, die den Stellenwert der Rechtschreibung für den Wechsel in die weiterführende Schulform untersuchen.

Der Bildungshintergrund wurde über die Bücheranzahl erhoben. Große Mittelwertsunterschiede sind zwischen wenigen Büchern (0-10 Stück), einer mittleren Anzahl verfügbarer Bücher (11-100 Stück) sowie vielen Büchern (ab 101 Stück) erkennbar (vgl. Abbildung 4.9 auf Seite 154). Die Kompetenzwerte auf Wort- und Teilkompetenzebene steigen im Verhältnis zu diesen drei gebildeten Kategorien der Bücheranzahl monoton an. Ab einem Besitz von über 100 Büchern erzielten die Schülerinnen und Schüler überdurchschnittliche Leistungen in allen Teilkompetenzen beider Tests. Gleichzeitig ist ab dieser Grenze der Einfluss auf die Höhe der Kompetenzwerte weniger deutlich erkennbar als unterhalb von 100 Büchern. Insgesamt betrachtet kann bezüglich des orthografischen Kompetenzstandes eine Kopplung mit der sozialen Herkunft beobachtet werden.

### Forschungsfrage VI

Die sechste und damit letzte Forschungsfrage erkundet die erreichten Leistungswerte in den Teilkompetenzen auf der Ebene einzelner Kinder sowie für die gesamte Stichprobe, um variierende Kompetenzmuster zu identifizieren und Profilverläufe aufzudecken.

Die explorative Analyse einzelner Kompetenzwerte von Kindern ließ intraindividuelle Unterschiede in den Ausprägungen der Teilkompetenzen des gutschrift-Tests und des SRT erkennen. Es wurden u. a. Schülerinnen und Schüler beschrieben, die eine oder mehrere Teilkompetenzen im Verhältnis zu den weiteren Teilkompetenzen sichtbar besser beherrschen (vgl. Tabelle 4.20 auf Seite 158). Die Prozent-korrekt-Werte verdeutlichten hier Differenzen von bis zu 69 Prozentpunkten (Fallbeispiel 3).

Um ganze Gruppen von Schülerinnen und Schülern mit unterscheidbaren Kompetenzprofilen zu ermitteln und weitere Informationen zu den Zusammenhängen zwischen den Teilkompetenzen der Rechtschreibtests zu erhalten, wurde eine latente Profilanalyse durchgeführt. Dafür erfolgte eine Berechnung von insgesamt 10 unterschiedlichen Klassenlösungen (vgl. Tabelle 4.21 auf Seite 162). Diese wurden u. a. anhand informationstheoretischer Maße und Likelihood-basierter Tests miteinander verglichen. Auf dieser Basis sowie inhaltlichen Überlegungen wurde eine 6-Klassenlösung für die nähere Betrachtung präferiert. In dieser Klassenlösung sind die Viertklässler in eine Ordnung gemäß ihrer erreichten Kompetenzwerte gruppiert (vgl. Abbildung 4.12 auf Seite 166). Die Kompetenzprofile unterscheiden sich demnach durch die Höhe des erzielten Niveaus in den Teilkompetenzen voneinander. Sie überschneiden sich an keiner Stelle. Dieses Ergebnis ist im Hinblick auf die zum Teil vorgefundenen hohen Korrelationen zwischen den Teilkompetenzen der Tests nicht erwartungswidrig.

Die Kinder in den Klassen, deren Kompetenzwerte um den Skalenmittelwert liegen, haben relativ ausgewogene Profilverläufe. Bei der leistungsstärksten und -schwächsten Klasse konnten die größten Variationen der relativen Kompetenzwerte ausgemacht werden. Auffällig erscheinen in der Gruppe der Rechtschreibkönner die vergleichsweise niedrigen Kompetenzwerte in den elementaren Teilkompetenzen (621 und 611 Punkte) und die zugleich relativ hohen Werte in den erweiterten Kompetenzen (656 und 691 Punkte) des gutschrift-Tests. In der Gruppe mit den leistungsschwächsten Kindern sind die im Verhältnis geringen Werte in den beiden Kernbereichen (287 und 301 Punkte) des SRT und der elementaren grammatischen Teilkompetenz (306 Punkte) von gutschrift unerwartet, da sie kleiner ausfallen als die Werte der erweiterten Teilkompetenzen (317 und 335 Punkte) und des Peripheriebereichs (316 Punkte). Dies könnte dafür sprechen, dass diese Schülerinnen und Schüler bisher kein grundlegendes Verständnis über den Aufbau und die Systematik der Schriftsprache entwickelt haben und sich Wortschreibungen eher durch Auswendiglernen, anstatt über Regeln und Einsichten, merken.

## 5.2 Diskussion der Ergebnisse und Forschungsdesiderate

Der Wert des Schreibens und des richtigen Schreibens ist für die Sicherstellung von Informationen und für die verständliche schriftliche Kommunikation unverkennbar. Die Orthografie, die ursprünglich als eine Erleichterung durch Normierung gedacht war, ist aber für nicht wenige Schriftlernende zu einem Problem geworden. Der Blickpunkt in der Diskussion um orthografisches Können sollte sich nicht ausschließlich auf *die eine* normgerechte Schreibung beziehen und Einzelfehler hervorheben. Stattdessen sollten die Beherrschung des Grundlegenden und eine Vermittlung der Regeln und der Systematik des Schriftsystems den Fokus bilden. Daher sollten Testinstrumente insbesondere dieses messen und Förderansätze hauptsächlich dieses berücksichtigen. Ein Rechtschreibkonzept, das die Orthografie mit möglichst vielen Regelhaftigkeiten und wenigen Ausnahmen beschreibt, steht dem Schriftsystem näher als ein Konzept, das dafür deutlich mehr Sonderregeln formuliert (Eisenberg, 1983, S. 54). Solch ein Konzept kann einem Schriftlernenden einen sinnvollen Zugang zur Welt des Schreibens eröffnen.

Mit gutschrift-diagnose und dem SRT stehen nun Testkonzepte zur Auswahl, die in unterschiedlichen Studien eingesetzt wurden bzw. werden, Rechtschreibung auf Basis eines theoretischen Hintergrundmodells differenziert betrachten und Förderansätze aufzeigen. Beide Modelle sehen Rechtschreibkompetenz nicht als globales Konstrukt an, sondern klassifizieren Fehl- und Richtigschreibungen auf einer qualitativen Ebene. Dabei werden Einteilungen zwischen leichteren und schwierigeren orthografischen Merkmalen vorgenommen. In gutschrift existieren diese in Form der elementaren und erweiterten Teilkompetenzen und im SRT kommen sie durch die Unterscheidung zwischen Kern- und Peripheriebereich zum Tragen. Diese Aufgliederung soll für Lehrpersonen und Kinder eine Orientierung beim Erwerbsprozess bieten. Sie konnte für beide Tests in dieser Arbeit über die empirisch ermittelten Schwierigkeitsgrade bestätigt werden. Ein zukünftiges und interessantes Forschungsgebiet stellt in diesem Kontext die Entwicklung von Kompetenzstufenmodellen für die beiden Rechtschreibtests dar. Diese sollten in Zusammenarbeit zwischen Fachdidaktik und Testentwicklung erstellt werden, damit die Bestimmung der Anforderung von Items zugleich auf Grundlage linguistisch fundierten sowie psychometrisch basierten Expertenwissens aufbaut.

Durch die gemeinsame neundimensionale Skalierung war es möglich, Statistiken (Korrelations- und Reliabilitätsanalysen) gleichzeitig über beide Tests zu ermitteln, sodass Aussagen zu den Teilkompetenzen der beiden Tests direkt aufeinander bezogen werden konnten. Für den Peripheriebereich konnte aufgezeigt werden, dass dieser sowohl unabhängig von den weiteren Teilkompetenzen des SRT als auch der des gutschrift-Tests ist. Die Eigenständigkeit dieses Bereichs ist damit aus psychometrischer Sicht bestätigt. Gleiche Befunde, in leicht abgeschwächter Form, liegen für das wortübergreifende Prinzip des SRT vor. Auch dieses erwies sich als vergleichsweise unabhängig zu den weiteren Teilkompetenzen des SRT und denen des gutschrift-Tests. Die hohen Zusammenhänge zwischen den Kernbereichen (phonographisch-silbisches und morphologisches Prinzip)

stellen gleichzeitig die Eigenständigkeit dieser Teilkompetenzen in Frage. Insgesamt betrachtet weist das Modell des SRT aber niedrigere korrelative Strukturen auf als das gutschrift-Kompetenzmodell. Die Befunde zum SRT können als gesichert gelten, da selbige Analysen mit den Daten der IGLU-Hauptuntersuchung nahezu identische Ergebnisse erbrachten. In einem nächsten Schritt wäre es interessant, den gutschrift-Test mit den Daten aus der Hauptuntersuchung zu analysieren, um die Dimensionalitätsanalysen anhand der größeren und repräsentativen Stichprobe durchzuführen.

Im Rahmen von Large-Scale-Assessments ist im Hinblick auf die Analyse von Kompetenzmodellen und Leistungswerten auch der Zusammenhang mit Hintergrundvariablen von Bedeutung. Für die Zusammenhangsanalysen in ZuRecht konnten die Variablen Geschlecht, Schulform und Bildungshintergrund genutzt werden. Es zeigten sich erwartungskonforme Ergebnisse: Mädchen erzielen bessere Testleistungen als Jungen, Rechtschreibkompetenz steht im Zusammenhang mit dem Übergang in die weiterführende Schulform und die soziale Schicht des Elternhauses moderiert die Leistungen. Gleichzeitig konnten differenzierte Ergebnisse für die Kompetenzmodelle gewonnen werden. Am auffälligsten war dabei die große Differenz auf Genderebene beim Peripheriebereich. Hier schneiden Mädchen bedeutsam besser ab als Jungen. Sie können sich demnach, gemäß dem SRT-Konzept, nicht regelhaft herleitbare Schreibungen besser merken bzw. haben Strategien dazu erschlossen.

Weiteren Einflussfaktoren auf die Rechtschreibkompetenz und den Rechtschreiberwerb kann im Rahmen groß angelegter Leistungsstudien nachgegangen werden. Die Daten aus gutschrift und SRT, die in der IGLU-Hauptuntersuchung eingesetzt worden sind, könnten mit den dort erhobenen Hintergrundinformationen verknüpft werden, da sie vielfältige Informationen zu Bedingungen des Lernprozesses über Schüler-, Eltern-, Lehrer- und Schulleitungsfragebögen sowie auch Angaben zum Rechtschreibunterricht beinhalten. Im Rahmen des Nationalen Bildungspanels findet ebenfalls eine Befragung von Schülerinnen und Schülern sowie Kontextpersonen statt. Die Studie hat ein längsschnittlich angelegtes Design und berücksichtigt Lerngelegenheiten in unterschiedlichen Lernumwelten (formale, non-formale und informelle) vom Kindes- bis ins hohe Erwachsenenalter (Blossfeld et al., 2011). In NEPS wurde der SRT für den Einsatz in der Sekundarstufe I weiterentwickelt (Strietholt et al., 2013, S. 573 f.). So liegen mittlerweile Testformen für mehrere Jahrgangsstufen vor. Die computergestützte Kodierung geht auf Arbeiten innerhalb dieser Studie zurück (Frahm, 2012). Damit bieten sowohl der gutschrift-Test als auch der SRT computerbasierte und standardisierte Auswertungen der Rechtschreibtests für unterschiedliche Jahrgangsstufen an.

Im Rahmen von ZuRecht wurde neben dem gutschrift-Test und dem SRT auch die HSP eingesetzt, die Strategien des Rechtschreiberwerbs auf entwicklungspsychologischer Grundlage unterscheidet. Die Daten der HSP wurden bis zum jetzigen Zeitpunkt noch nicht in differenzierter Weise ausgewertet. Ein zukünftiges Forschungsvorhaben wäre, analog zu den hier erfolgten Analysen, eine transparente und vollständige Auswertung der HSP-Strategien und eine Einordung der Ergebnisse in diese Studie sowie ein Vergleich der messtheoretischen Werte anzustreben. Dies kann weitere Hinweise auf das Konstrukt der Rechtschreibkompetenz liefern.

Mit Hilfe von Studien wie ZuRecht, die die fachtheoretische Modellierung der Rechtschreibkompetenz übernehmen, um sie einer empirisch vergleichenden Überprüfung zu unterziehen, können differenzierte und gesicherte Aussagen zum latenten Konstrukt der Orthografie mit dem Ziel formuliert werden, alle Schülerinnen und Schüler zukünftig besser in ihrem Lernprozess unterstützen zu können und so Wege aus der Bildungsbenachteiligung zu finden. In Anbetracht der Unterschiedlichkeit der den Rechtschreibtests zugrundeliegenden theoretischen Konzepte kann und soll an dieser Stelle aber keine eindeutige Aussage für oder gegen einen Test bzw. ein Modell getroffen werden. Es können ausschließlich Vor- und Nachteile eines Tests abgewogen werden, wobei hier stets die individuellen Bedürfnisse und Voraussetzungen des Lerners Berücksichtigung finden sollten, da dieser bei allen erziehungswissenschaftlichen Fragestellungen stets den Mittelpunkt bilden muss.

# Literaturverzeichnis


Adams, R. (2002). Scaling PISA Cognitive Data. In R. Adams & M. Wu (Hrsg.), *PISA 2000 Technical Report* (Kap. 9, S. 99–108). Zugriff am 28.04.2011 unter http://www.oecd.org/pisa/pisaproducts/33688233.pdf. Organisation for Economic Co-Operation and Development. (Siehe S. 97, 98).

Adams, R. (2005). Reliability as a Measurement Design Effect. *Studies in Educational Evaluation*, *31*, 162–172. (Siehe S. 138).

Adams, R. & Carstensen, C. H. (2002). Scaling Outcomes. In R. Adams & M. Wu (Hrsg.), *PISA 2000 Technical Report* (Kap. 13, S. 149–162). Zugriff am 28.04.2011 unter http://www.oecd.org/pisa/pisaproducts/33688233.pdf. Organisation for Economic Co-Operation and Development. (Siehe S. 122, 123, 138).

Adams, R., Wilson, M. & Wang, W.-c. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, *21*(1), 1–23. (Siehe S. 90, 91, 135).

Adams, R. & Wu, M. (2007). The Mixed-Coefficients Multinominal Logit Model: A Generalized Form of the Rasch Model. In M. von Davier & C. H. Carstensen (Hrsg.), *Multivariate and Mixture Distribution Rasch Models. Extensions and Applications* (S. 57–76). Statistics for Social and Behavioral Science. New York, Berlin: Springer. (Siehe S. 90).

Altmann, H. & Ziegenhain, U. (2007). *Phonetik, Phonologie und Graphemik fürs Examen* (2. Auflage). Göttingen: Vandenhoeck & Ruprecht. (Siehe S. 33, 44).

Asparouhov, T. & Muthén, B. (2012). *Using Mplus TECH11 and TECH14 to test the number of latent classes*. Mplus Web Notes: No. 14. Zugriff am 14.08.2013 unter https://www.statmodel.com/examples/webnotes/webnote14.pdf.

Augst, G. & Dehn, M. (1998). *Rechtschreibung und Rechtschreibunterricht. Können – Lehren – Lernen. Eine Einführung für Studierende und Lehrende aller Schulformen*. Stuttgart, Düsseldorf, Leipzig: Klett Verlag.

Bacher, J., Pöge, A. & Wenzig, K. (2010). *Clusteranalyse. Eine anwendungsorientierte Einführung in Klassifikationsverfahren* (3., ergänzte, vollständig überarbeitete und

neu gestaltete Auflage). München: Oldenbourg Wissenschaftsverlag. (Siehe S. 100, 105, 106).

Baker, F. B. (2001). *The Basics of Item Response Theory* (2. Aufl.). ERIC Clearinghouse on Assessment und Evaluation. (Siehe S. 82, 91).

Balhorn, H. (1993). Diagnose und förderung in der rechtschreibung. *Diskussion Deutsch*, *132*(24), 307–317.

Beck, B., Bundt, S. & Gomolka, J. (2008). Ziele und Anlage der DESI-Studie. In DESI-Konsortium (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (S. 11–25). Weinheim, Basel: Beltz. (Siehe S. 1).

Beck, B., Thomé, G. & Thomé, D. (2009). Schwache Rechtschreiber müssen keine schwachen Leser sein und umgekehrt. Ergebnisse aus der Schulleistungsstudie DESI. In R. Valtin & B. Hofmann (Hrsg.), *Kompetenzmodelle der Orthographie. Empirische Befunde und förderdiagnostische Möglichkeiten* (Bd. 10, S. 40–47). dgLs Beiträge. Berlin: Deutsche Gesellschaft für Lesen und Schreiben. (Siehe S. 2, 152).

Bereiter, C. (1980). Development in writing. In L. W. Gregg & E. R. Steinberg (Hrsg.), *Cognitive Processes in Writing* (S. 73–93). Hillsdale: Lawrence Erlbaum.

Blatt, I. (2006). Am Dehnungs-h zweifeln, aber nicht verzweifeln. Kinder erforschen, üben und festigen das Dehnungs-h. *Praxis Deutsch*, *33*(198), 28–35. (Siehe S. 63).

Blatt, I. (2007a). Begleitheft für Lehrerinnen und Lehrer zum Lernheft Rechtschreiben – Grundlagen (Klasse 4-6). Lernheft Schule Nr. 1. Manuskriptdruck.

Blatt, I. (2007b). Lernheft Rechtschreiben – Grundlagen (Klasse 4-6). Lernheft Schule Nr. 1. Manuskriptdruck.

Blatt, I. (2007c). Lösungen für das Lernheft Rechtschreiben – Grundlagen (Klasse 4-6). Lösungsheft Schule Nr. 1. Manuskriptdruck.

Blatt, I. (2010). Sprachsystematische Rechtschreibdidaktik: Konzept, Materialien, Tests. In U. Bredel, A. Müller & G. Hinney (Hrsg.), *Schriftsystem und Schrifterwerb: linguistisch – didaktisch – empirisch* (S. 101–132). Berlin, New York: De Gruyter. (Siehe S. 51–53, 58, 60, 63, 65, 66).

Blatt, I. (2011). Wie lässt sich Textqualität ermitteln und lernförderlich zurückmelden? In J. Berning (Hrsg.), *Textwissen und Schreibbewusstsein: Beiträge aus Forschung und Praxis* (S. 89–114). Münster: LIT 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. (Siehe S. 52–55, 64, 66).

Blatt, I., Frahm, S. & Jarsinski, S. (2012). *NEPS Technical Report for Orthography -– Scaling Results of Starting Cohort 3 in Fifth Grade*. Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel. NEPS Working Paper No. 21. Zugriff am 11.12.2012 unter http://www.neps-data.de/Portals/0/Working%20Papers/WP_XXI.pdf.

Blatt, I. & Jarsinski, S. (2009). Auswertung nach der Sprachsystematischen Rechtschreibdiagnose. In R. Valtin & B. Hofmann (Hrsg.), *Kompetenzmodelle der Orthographie. Empirische Befunde und förderdiagnostische Möglichkeiten* (Bd. 10, S. 91–112). dgLs Beiträge. Berlin: Deutsche Gesellschaft für Lesen und Schreiben. (Siehe S. 53, 73).

Blatt, I., Müller, A. & Voss, A. (2010). Schriftstruktur als Lesehilfe. Konzeption und Ergebnisse eines Hamburger Leseförderprojekts in Klasse 5 (HeLp). In U. Bredel, A. Müller & G. Hinney (Hrsg.), *Schriftsystem und Schrifterwerb: linguistisch – didaktisch – empirisch* (S. 171–202). Berlin, New York: De Gruyter.

Blatt, I. & Voss, A. (2006). Rechtschreibtest Klasse 4 aus der IGLU 2006-Voruntersuchung. In W. Bos & S. Hornberg (Hrsg.), *Dokumentation der Fachtagung zu Rechtschreibtests IGLU 2006* (S. 24–29). Dortmund: Internes Papier am Institut für Schulentwicklungsforschung.

Blatt, I., Voss, A., Kowalski, K. & Jarsinski, S. (2011). Messung von Rechtschreibleistung und empirische Kompetenzmodellierung. In U. Bredel & T. Reißig (Hrsg.), *Weiterführender Orthographieerwerb* (Bd. 5, S. 226–256). Deutschunterricht in Theorie und Praxis (DTP). Baltmannsweiler: Schneider Hohengehren. (Siehe S. 43, 51–56, 59, 66, 80, 130).

Blossfeld, H.-P., von Maurice, J. & Schneider, T. (2011). The National Educational Panel Study: need, main features, and research potential. *Zeitschrift für Erziehungswissenschaft*, *14*(Sonderheft 14), 5–18. (Siehe S. 4, 178).

Böhme, K. & Bremerich-Vos, A. (2009). Diagnostik der Rechtschreibkompetenz in der Grundschule -– Konstruktprüfung mittels Fehler- und Dimensionalitätsanalysen. 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, Basel: Beltz. (Siehe S. 18, 22–24, 66, 150).

Böhme, K. & Bremerich-Vos, A. (2012). Beschreibung der im Fach Deutsch untersuchten Kompetenzen. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 19–33). Münster, New York, München, Berlin: Waxmann. (Siehe S. 25, 26).

Böhme, K., Richter, D., Stanat, P., Pant, H. A. & Köller, O. (2012). Die länderübergreifenden Bildungsstandards in Deutschland. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 11–18). Münster, New York, München, Berlin: Waxmann. (Siehe S. 25).

Böhme, K. & Roppelt, A. (2012). Geschlechtsbezogene Disparitäten. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 173–190). Münster, New York, München, Berlin: Waxmann. (Siehe S. 150).

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (2. Aufl.). Mahwah, New Jersey: Lawrence Erlbaum. (Siehe S. 84, 85, 87, 89, 97, 99, 100, 119, 121).

Bonsen, M., Büchter, A. & van Ophysen, S. (2004). Im Fokus: Leistung. Zentrale Aspekte der Schulleistungsforschung und ihre Bedeutung für die Schulentwicklung. In H. G. Holtappels, H. Pfeiffer, H.-G. Rolff, R. Schulz-Zander & K. Klemm (Hrsg.), *Jahr-*

*buch der Schulentwicklung. Daten, Beispiele und Perspektiven* (Bd. 13, S. 187–224). Jahrbuch der Schulentwicklung. Weinheim, München: Juventa. (Siehe S. 146).

Borchert, J., Knopf-Jerchow, H. & Dahbashi, A. (1991). *Testdiagnostische Verfahren in Vor-, Sonder- und Regelschulen. Ein kritisches Handbuch für Praktiker*. Heidelberg: Roland Asanger Verlag.

Borg, I. & Staufenbiel, T. (2007). *Lehrbuch Theorien und Methoden der Skalierung* (4., vollständig überarbeitete und erweiterte Auflage). Bern: Verlag Hans Huber. (Siehe S. 23, 80, 81, 86, 87, 91, 93–95).

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4., überarbeitete Auflage. Mit 156 Abbildungen und 87 Tabellen). Heidelberg: Springer. (Siehe S. 17, 81, 83, 96–98, 139, 146, 147).

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7., vollständig überarbeitete und erweiterte Auflage. Mit 70 Abbildungen und 163 Tabellen). Berlin, Heidelberg, New York: Springer. (Siehe S. 100).

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. 91–136). Münster, New York, München, Berlin: Waxmann. (Siehe S. 150, 155).

Bos, W., Brose, U., Bundt, S., Gröhlich, C., Hugk, N., Janke, N., … Voss, A. (2006). Anlage und Durchführung der Studie ‚Kompetenzen und Einstellungen von Schülerinnen und Schülern -– Jahrgangsstufe 4 (KESS 4)‘. In W. Bos & M. Pietsch (Hrsg.), *KESS 4 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen* (Bd. 1, S. 9–32). Hanse – Hamburger Schriften zur Qualität im Bildungswesen. Münster, New York, München, Berlin: Waxmann. (Siehe S. 10).

Bos, W., Gröhlich, C., Guill, K., Scharenberg, K. & Wendt, H. (2010). Ziele und Anlage der Studie KESS 8. In W. Bos & C. Gröhlich (Hrsg.), *KESS 8 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 8* (Bd. 6, S. 9–20). Hanse -– Hamburger Schriften zur Qualität im Bildungswesen. Münster, New York, München, Berlin: Waxmann. (Siehe S. 10).

Bos, W., Hornberg, S., Arnold, K.-H., Faust, G., Fried, L., Lankes, E.-M., … Valtin, R. (2008). IGLU 2006: Eine Schulleistungsstudie der IEA und ihre nationale Erweiterung. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, … R. Valtin (Hrsg.), *IGLU-E 2006. Die Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich* (S. 9–16). Münster, New York, München, Berlin: Waxmann. (Siehe S. 3).

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Valtin, R., Voss, A. & Walther, G. (Hrsg.). (2005). *IGLU. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente*. Münster, New York, München, Berlin: Waxmann. (Siehe S. 66).

Bos, W., Lankes, E.-M., Schwippert, K., Valtin, R., Voss, A., Badel, I. & Plaßmeier, N. (2003). Lesekompetenzen deutscher Grundschülerinnen und Grundschüler am Ende der vierten Jahrgangsstufe im internationalen Vergleich. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin & G. Walther (Hrsg.), *Erste Ergebnisse aus*

*IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich* (S. 69–142). Münster, New York, München, Berlin: Waxmann. (Siehe S. 29, 123).

Bos, W., Pietsch, M., Poerschke, J. & Vieluf, U. (2006). Zusammenfassung wichtiger Ergebnisse zu Kompetenzen und Einstellungen von Hamburger Schülerinnen und Schülern. In W. Bos & M. Pietsch (Hrsg.), *KESS 4 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen* (Bd. 1, S. 1–8). Hanse – Hamburger Schriften zur Qualität im Bildungswesen. Münster, New York, München, Berlin: Waxmann.

Bos, W., Pietsch, M. & Stubbe, T. C. (2006). Regionale, nationale und internationale Einordnung der Lesekompetenz und weiterer Schulleistungsergebnisse Hamburger Kinder am Ende der Grundschulzeit. In W. Bos & M. Pietsch (Hrsg.), *KESS 4 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen* (Bd. 1, S. 57–86). Hanse – Hamburger Schriften zur Qualität im Bildungswesen. Münster, New York, München, Berlin: Waxmann. (Siehe S. 16).

Bos, W., Schwippert, K. & Stubbe, T. C. (2007). Die Kopplung von sozialer Herkunft und Schülerleistung im internationalen Vergleich. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, . . . R. Valtin (Hrsg.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 225–248). Münster, New York, München, Berlin: Waxmann. (Siehe S. 153, 155).

Bos, W., Sereni, S. & Stubbe, T. C. (2008). IGLU Belgien im Kontext von PIRLS 2006. In W. Bos, S. Sereni & T. C. Stubbe (Hrsg.), *IGLU Belgien. Lese- und Orthografiekompetenzen von Grundschulkindern in der Deutschsprachigen Gemeinschaft* (S. 11–18). Münster, New York, München, Berlin: Waxmann. (Siehe S. 41).

Bos, W. & Tarnai, C. (Hrsg.). (1989). *Angewandte Inhaltsanalyse in Empirischer Pädagogik und Psychologie*. Münster, New York: Waxmann. (Siehe S. 76).

Bos, W. & Tarnai, C. (Hrsg.). (1996). *Computerunterstützte Inhaltsanalyse in den empirischen Sozialwissenschaften: Theorie, Anwendung, Software*. Münster, New York: Waxmann.

Bos, W., Valtin, R., Hornberg, S., Buddeberg, I., Goy, M. & Voss, A. (2007). Internationaler Vergleich 2006: Lesekompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, . . . R. Valtin (Hrsg.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 109–160). Münster, New York, München, Berlin: Waxmann. (Siehe S. 123).

Bos, W., Valtin, R., Voss, A., Hornberg, S. & Lankes, E.-M. (2007). Konzepte der Lesekompetenz in IGLU 2006. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, . . . R. Valtin (Hrsg.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 81–108). Münster, New York, München, Berlin: Waxmann. (Siehe S. 122, 123).

Bos, W., Wendt, H., Ünlü, A., Valtin, R., Euen, B., Kasper, D. & Tarelli, I. (2012a). Leistungsprofile von Viertklässlerinnen und Viertklässlern in Deutschland. In W. Bos, I. Tarelli, A. Bremerich-Vos & K. Schwippert (Hrsg.), *IGLU 2011. Lesekompetenzen*

*von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 227–259). Münster, New York, München, Berlin: Waxmann. (Siehe S. 107–109, 124, 166).

Bos, W., Wendt, H., Ünlü, A., Valtin, R., Euen, B., Kasper, D. & Tarelli, I. (2012b). Leistungsprofile von Viertklässlerinnen und Viertklässlern in Deutschland. In W. Bos, H. Wendt, O. Köller & C. Selter (Hrsg.), *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 269–302). Münster, New York, München, Berlin: Waxmann. (Siehe S. 107–109).

Bredel, U. (2006). Orthographische Zweifelsfälle. *Praxis Deutsch*, *198*, 6–15.

Bredel, U. & Günther, H. (2006). Orthographietheorie und Rechtschreibunterricht. In U. Bredel & H. Günther (Hrsg.), *Orthographietheorie und Rechtschreibunterricht* (Bd. 509, S. 197–215). Linguistische Arbeiten. Tübingen: Niemeyer.

Bremerich-Vos, A. (1996). Aspekte des Schriftspracherwerbs. Stufentheorien, das „Neue“ und die Lehrer-Schüler-Interaktion. In A. Peyer & P. R. Portmann (Hrsg.), *Norm, Moral und Didaktik – Die Linguistik und ihre Schmuddelkinder. Eine Aufforderung zur Diskussion* (S. 267–290). Tübingen: Max Niemeyer Verlag.

Bremerich-Vos, A. (2004). Rechtschreibstandards, Kompetenzstufen und IGLU — einige Anmerkungen. In A. Bremerich-Vos, C. Löffler & K.-L. Herné (Hrsg.), *Neue Beiträge zur Rechtschreibtheorie und -didaktik* (S. 85–104). Freiburg: Fillibach. (Siehe S. 32, 35, 58, 59, 63, 64).

Bremerich-Vos, A. (2010). Modellierung von Aspekten sprachlich-kultureller Kompetenz. Anmerkungen zu den Projektberichten. *Zeitschrift für Pädagogik*, (56), 199–203.

Bremerich-Vos, A. (2011). Die Bildungsstandards Deutsch. In A. Bremerich-Vos, D. Granzer, U. Behrens & O. Köller (Hrsg.), *Bildungsstandards für die Grundschule: Deutsch konkret. Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen* (3. Aufl., S. 14–42). Cornelsen.

Bremerich-Vos, A., Böhme, K., Krelle, M., Weirich, S. & Köller, O. (2012). Kompetenzstufenmodelle im Fach Deutsch. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 56–71). Münster, New York, München, Berlin: Waxmann. (Siehe S. 26–28).

Bremerich-Vos, A., Böhme, K. & Robitzsch, A. (2009). Von Bildungsstandards zu ihrer überprüfung: Grundlagen der Item- und Testentwicklung. 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. 198–218). Weinheim, Basel: Beltz.

Briggs, D. C. & Wilson, M. (2003). An Introduction to Multidimensional Measurement using Rasch Models. *Journal of Applied Measurement*, *4*(1), 87–100.

Brügelmann, H. (2003). Rechtschreiben am Ende der Grundschulzeit: 1991-2001. NRW-KIDS 2001 und der Schreibvergleich Bundesrepublik-DDR. In A. Panagiotopoulou & H. Brügelmann (Hrsg.), *Grundschulpädagogik meets Kindheitsforschung. Zum Wechselverhältnis von schulischem Lernen und außerschulischen Erfahrungen im Grundschulalter* (Bd. 7, S. 173–178). Jahrbuch Grundschulforschung. Opladen: Leske + Budrich.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte und erw. Aufl.). München: Pearson Studium. (Siehe S. 94, 95, 104, 138).

Bühner, M. & Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler*. München: Pearson Studium. (Siehe S. 88).

Burnham, K. P. & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304. (Siehe S. 92, 95).

Butt, M. & Eisenberg, P. (1990). Schreibsilbe und Sprechsilbe. In C. Stetter (Hrsg.), *Zu einer Theorie der Orthographie. Interdisziplinäre Aspekte gegenwärtiger Schrift- und Orthographieforschung* (Bd. 99, S. 33–64). Reihe Germanistische Linguistik. Tübingen: Max Niemeyer Verlag. (Siehe S. 43, 45–47).

Carstensen, C. H. (2000). *Mehrdimensionale Testmodelle mit Anwendungen aus der pädagogisch-psychologischen Diagnostik*. Kiel: IPN. (Siehe S. 80, 81, 89, 93, 94, 99).

Celeux, G. & Soromenho, G. (1996). An Entropy Criterion for Assessing the Number of Clusters in a Mixture Model. *Journal of Classification*, (13), 195–212. (Siehe S. 107).

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2. Aufl.). Hillsdale, New Jersey: Lawrence Erlbaum. (Siehe S. 146).

Collins, L. M. & Lanza, S. T. (2010). *Latent Class and Latent Transition Analysis. With Applications in the Social, Behavioral, and Health Sciences*. Hoboken, New Jersey: Wiley. (Siehe S. 100–102, 105, 107, 161).

de Gruijter, D. N. M. & van der Kamp, L. J. T. (2008). *Statistical Test Theory for the Behavioral Sciences*. Statistics in the Social and Behavioral Sciences Series. Baco Raton, London, New York: Chapmann & Hall/CRC. (Siehe S. 81, 84, 86, 91, 94).

DeMars, C. (2010). *Item Response Theory. Understanding Statistics: Measurement*. Series in Understanding Statistics. Oxford, New York: Oxford University Press. (Siehe S. 126).

Dias, J. G. & Vermunt, J. K. (2006). Bootstrap methods for measuring classification uncertainty in latent class analysis. In A. Rizzi & M. Vichi (Hrsg.), *Compstat 2006 - Proceedings in Computational Statistics* (S. 31–42). Heidelberg: Physica Verlag. (Siehe S. 107).

Drechsel, B. & Artelt, C. (2007). Lesekompetenz. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Hrsg.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 225–248). Münster, New York, München, Berlin: Waxmann. (Siehe S. 150).

Eisenberg, P. (1983). Orthografie und Schriftsystem. In K. B. Günther & H. Günther (Hrsg.), *Schrift, Schreiben, Schriftlichkeit. Arbeiten zur Struktur, Funktion und Entwicklung schriftlicher Sprache* (Bd. 49, S. 41–68). Reihe Germanistische Linguistik. Tübingen: Max Niemeyer Verlag. (Siehe S. 43, 177).

Eisenberg, P. (1988). Die Grapheme des Deutschen und ihre Beziehung zu den Phonemen. In J. Baumert, K. B. Günther & U. Knoop (Hrsg.), *Aspekte von Schrift und Schriftlichkeit* (S. 139–154). Hildesheim, Zürich, New York: Georg Olms Verlag. (Siehe S. 44).

Eisenberg, P. (1989). Die Schreibsilbe im Deutschen. In P. Eisenberg & H. Günther (Hrsg.), *Schriftsystem und Orthographie* (Bd. 97, S. 57–84). Reihe Germanistische Linguistik. Tübingen: Max Niemeyer Verlag. (Siehe S. 43–47, 60, 62).

Eisenberg, P. (1990). Die Sprache und die Schrift. Warum es so schwierig ist, unsere Orthographie zu reformieren. *Praxis Deutsch*, *17*(103), 4–7. (Siehe S. 60).

Eisenberg, P. (1993). Linguistische Fundierung orthographischer Regeln. Umrisse einer Wortgraphematik des Deutschen. In J. Baurmann, H. Günther & U. Knoop (Hrsg.), *Homo Scribens. Perspektiven der Schriftlichkeitsforschung* (Bd. 134, S. 67–94). Reihe Germanistische Linguistik. Tübingen: Max Niemeyer Verlag.

Eisenberg, P. (2006a). *Das Wort* (3., durchgesehene Auflage). Grundriss der deutschen Grammatik. Stuttgart, Weimar: Metzler. (Siehe S. 44–49, 60, 63, 67).

Eisenberg, P. (2006b). *Der Satz* (3., durchgesehene Auflage). Grundriss der deutschen Grammatik. Stuttgart, Weimar: Metzler.

Eisenberg, P. (2006c). Phonem und Graphem. In D. Dudenredaktion (Hrsg.), *Duden. Die Grammatik* (7., völlig neu erarbeitete und erweiterte Auflage, Bd. 4, S. 19–94). Manheim, Leipzig, Wien, Zürich: Dudenverlag. (Siehe S. 33, 43–48, 60, 62, 68).

Eisenberg, P. (2011). Grundlagen der deutschen Wortschreibung. In U. Bredel & T. Reißig (Hrsg.), *Weiterführender Orthographieerwerb* (Bd. 5, S. 83–95). Deutschunterricht in Theorie und Praxis (DTP). Baltmannsweiler: Schneider Hohengehren. (Siehe S. 46–48, 68).

Eisenberg, P. & Feilke, H. (2001). Rechtschreiben erforschen. *Praxis Deutsch*, (170), 6–15. (Siehe S. 49, 62).

Eisenberg, P. & Fuhrhop, N. (2007). Schulorthographie und Graphematik. *Zeitschrift für Sprachwissenschaft*, (26), 15–41. (Siehe S. 43).

Eisenberg, P. & Menzel, W. (1995). Grammatik-Werkstatt. *Praxis Deutsch*, (129), 14–27.

Eisenberg, P., Ramers, K. H. & Vater, H. (Hrsg.). (1992). *Silbenphonologie des Deutschen*. Studien zur deutschen Grammatik. Tübingen: Narr.

Eisenberg, P., Spitta, G. & Voigt, G. (1994). Schreiben: Rechtschreiben. *Praxis Deutsch*, (124), 14–25. (Siehe S. 45–49, 63).

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Multivariate Applications books series. Mahwah, New Yersey: Lawrence Erlbaum. (Siehe S. 5, 81–83, 85, 88, 91, 92, 99, 126, 156).

Erdfelder, E., Faul, F., Buchner, A. & Cüpper, L. (2010). Effektgröße und Teststärke. In H. Holling & B. Schmitz (Hrsg.), *Handbuch Statistik, Methoden und Evaluation* (Bd. 13, S. 358–369). Handbuch der Psychologie. Göttingen, Bern, Wien, Paris, Oxford, Prag, Toronto, Cambridge, Amsterdam, Kopenhagen, Stockholm: Hogrefe. (Siehe S. 146).

Fischer, G. H. (1968). *Psychologische Testtheorie*. Bern, Stuttgart: Hans Huber. (Siehe S. 81, 87, 88).

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendungen*. Bern, Stuttgart, Wien: Hans Huber. (Siehe S. 5, 79, 81, 82, 84, 86, 88, 93, 94).

Formann, A. K. (2010). Latent-Class-Analyse. In H. Holling & B. Schmitz (Hrsg.), *Handbuch Statistik, Methoden und Evaluation* (Bd. 13, S. 556–561). Handbuch der Psycho-

logie. Göttingen, Bern, Wien, Paris, Oxford, Prag, Toronto, Cambridge, Amsterdam, Kopenhagen, Stockholm: Hogrefe. (Siehe S. 105, 106).

Foy, P., Galia, J. & Li, I. (2007). Scaling the PIRLS 2006 Reading Assessment Data. In M. O. Martin, I. V. S. Mullis & A. M. Kennedy (Hrsg.), *PIRLS 2006 Technical Report* (S. 149–172). Zugriff am 01.08.2013 unter http://timssandpirls.bc.edu/PDF/p06_technical_report.pdf. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College. (Siehe S. 127, 145).

Frahm, S. (2012). *Computerbasierte Testung der Rechtschreibleistung in Klasse fünf – eine empirische Studie zu Mode-Effekten im Kontext des Nationalen Bildungspanels*. Berlin: Logos Verlag. (Siehe S. 53, 55, 77, 178).

Frahm, S., Goy, M., Kowalski, K., Sixt, M., Strietholt, R., Blatt, I., … Kanders, M. (2011). Transition and development from lower secondary to upper secondary school. *Zeitschrift für Erziehungswissenschaft*, (Sonderheft 14), 217–232. (Siehe S. 77).

Frey, A., Carstensen, C. H., Walter, O., Rönnebeck, S. & Gomolka, J. (2008). Methodische Grundlagen des Ländervergleichs. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Hrsg.), *PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich* (S. 375–398). Münster, New York, München, Berlin: Waxmann. (Siehe S. 145).

Frith, U. (1985). Beneath the Surface of Developmental Dyslexia. In K. E. Patterson, J. C. Marshall & M. Coltheart (Hrsg.), *Surface Dyslexia. Neuropsychological and Cognitive Studies of Phonological Reading* (S. 301–330). London: Lawrence Erlbaum. (Siehe S. 12).

Frith, U. (1986). Psychologische Aspekte des orthographischen Wissens. Entwicklung und Entwicklungsstörung. In G. Augst (Hrsg.), *New Trends in Graphemics and Orthography* (S. 218–233). Berlin: Walter de Gruyter. (Siehe S. 12).

Fuhrhop, N. (2006). *Orthografie* (2. aktualisierte Auflage). Kurze Einführungen in die germanistische Linguistik. Heidelberg: Universitätsverlag Winter. (Siehe S. 44–49).

Fuhrhop, N. (2008). Das graphematische Wort (im Deutschen): Eine erste Annäherung. *Zeitschrift für Sprachwissenschaft*, *27*(2), 189–228.

Furr, R. M. & Bacharach, V. R. (2008). *Psychometrics: An Introduction*. Los Angeles, London, New Delhi, Singapore: Sage. (Siehe S. 82, 84, 126).

Gautschi, T. (2010). Maximum-Likelihood Schätztheorie. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 205–238). Wiesbaden: VS Verlag. (Siehe S. 91, 92).

Geiser, C. (2011). *Datenanalyse mit Mplus. Eine anwendungsorientierte Einführung* (2., durchgesehene Auflage). Wiesbaden: VS Verlag. (Siehe S. 101, 106, 107, 164).

Geiser, C. & Eid, M. (2010). Item-Response-Theorie. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 311–332). Wiesbaden: VS Verlag. (Siehe S. 83, 84).

Ghanbari, S. A. (2002). *Einführung in Die Statistik für Sozial- und Erziehungswissenschaftler*. Berlin, Heidelberg, New York, Barcelona, Hongkong, London, Mailand, Paris, Tokio: Springer. (Siehe S. 112).

Gniewosz, B. (2011a). Kompetenzentwicklung. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Empirische Bildungsforschung. Gegenstandsberichte* (S. 57–67). Wiesbaden: VS Verlag.

Gniewosz, B. (2011b). Testverfahren. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Empirische Bildungsforschung. Strukturen und Methoden* (S. 67–76). VS Verlag. (Siehe S. 79, 80).

Gollwitzer, M. (2008). Latent-Class-Analysis. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 279–306). Mit 79 Abbildungen und 43 Tabellen. Heidelberg: Springer. (Siehe S. 91, 92, 95, 96, 101, 103, 105, 106).

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231. Zugriff am 03.09.2013 unter http://www.statmodel.com/bmuthen/ED231e/RelatedArticles/Goodman.pdf. (Siehe S. 105).

Granzer, D. (2009). Von Bildungsstandards zu ihrer überprüfung: Grundlagen der Item- und Testentwicklung. 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. 21–30). Weinheim, Basel: Beltz.

Granzer, D., Böhme, K. & Köller, O. (2008). Kompetenzmodelle und Aufgabenentwicklung für die standadisierte Leistungsmessung im Fach Deutsch. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Lernstandsbestimmungen im Fach Deutsch. Gute Aufgaben für den Unterricht* (S. 10–28). Weinheim, Basel: Beltz.

Granzer, D., Köller, O., Bremerich-Vos, A., van den Heuvel-Panhuizen, M., Reiss, K. & Walther, G. (2009). Vorwort. 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. 7–9). Weinheim, Basel: Beltz. (Siehe S. 18).

Günther, H. (1993). Erziehung zur Schriftlichkeit. In P. Eisenberg & P. Klotz (Hrsg.), *Sprache gebrauchen – Sprachwissen erwerben* (S. 85–96). Stuttgart, Düsseldorf, Berlin, Leipzig: Ernst Klett Schulbuchverlag.

Günther, H. & Nünke, E. (2005). Warum das Kleine groß geschrieben wird, wie man das lernt und wie man das lehrt. *Kölner Beiträge zur Sprachdidaktik*, (1). Zugriff am 16.04.2014 unter http://www.koebes.uni-koeln.de/guenther_nuenke.pdf. (Siehe S. 49).

Haag, N. & Roppelt, A. (2012). Der Ländervergleich im Fach Mathematik. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 117–127). Münster, New York, München, Berlin: Waxmann.

Hartig, J. (2004). Methoden zur Bildung von Kompetenzstufenmodellen. In H. Moosbrugger, D. Frank & W. Rauch (Hrsg.), *Qualitätssicherung im Bildungswesen* (Bd. 3, S. 74–93). Frankfurt am Main: Institut für Psychologie der Johann Wolfgang Goethe-Universität.

Hartig, J. (2008a). Kompetenzen als Ergebnisse von Bildungsprozessen. In N. Jude, J. Hartig & E. Klieme (Hrsg.), *Kompetenzerfassung in pädagogischen Handlungsfeldern. Theorien, Konzepte und Methoden* (Bd. 26, S. 15–25). Bildungsforschung. Bonn, Berlin: Bundesministerium für Bildung und Forschung.

Hartig, J. (2008b). Kompetenzen als Ergebnisse von Bildungsprozessen. In *Kompetenzerfassung in pädagogischen Handlungsfeldern* (Bd. 26, S. 13–24). Bildungsforschung. Bonn, Berlin: Bundesministerium für Bildung und Forschung.

Hartig, J. (2008c). Psychometric Models for the Assessment of Competencies. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of Competencies in Educational Contexts* (S. 69–90). New York: Hogrefe & Huber.

Hartig, J. & Höhler, J. (2008). Representation of Competencies in Multidimensional IRT Models with Within-Item and Between-Item Multidimensionality. *Journal of Psychology*, *216*(2), 89–101.

Hartig, J. & Höhler, J. (2010). Modellierung von Kompetenzen mit mehrdimensionalen IRT-Modellen. *Zeitschrift für Pädagogik*, (56), 189–198. (Siehe S. 89).

Hartig, J. & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In J. Hartig & E. Klieme (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung* (Bd. 20, S. 17–36). Bildungsforschung. Bonn, Berlin: Bundesministerium für Bildung und Forschung.

Hartig, J., Jude, N. & Wagner, W. (2008). Methodische Grundlagen der Messung und Erklärung sprachlicher Kompetenzen. In DESI-Konsortium (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (S. 34–54). Weinheim, Basel: Beltz.

Hartig, J. & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127–143). Berlin: Springer.

Hasselhorn, M., Marx, H. & Schneider, W. (2008). Diagnose der orthografischen Kompetenz — von der HSP zur DSP. In W. Schneider, H. Marx & M. Hasselhorn (Hrsg.), *Diagnostik von Rechtschreibleistungen und -kompetenz* (Bd. 6, S. 1–6). Jahrbuch der pädagogisch-psychologischen Diagnostik. Tests und Trends. Göttingen, Bern, Wien, Paris, Oxford, Prag, Toronto, Cambridge, Amsterdam, Kopenhagen: Hogrefe. (Siehe S. 2, 3).

Helmke, A. & Weinert, F. E. (1997). Die Münchener Grundschulstudie SCHOLASTIK: Wissenschaftliche Grundlagen, Zielsetzungen, Realisierungsbedingungen und Ergebnisperspektiven. In F. E. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S. 1–12). Weinheim: Beltz.

Herné, K.-L. (1993). Der schmusige Elecktriger im Omibus tabde in einen fett Topf. überlegungen zu einer förderdiagnostischen Rechtschreibfehler-Klassifikation. *Diskussion Deutsch*, *24*(132), 318–328. (Siehe S. 20, 21, 64, 65, 68).

Herné, K.-L. (2003). Rechtschreibtests. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch* (2., durchgesehene Auflage, Bd. 2, S. 883–897). Paderborn, München, Wien, Zürich: Ferdinand Schöningh. (Siehe S. 13, 20, 68).

Herné, K.-L. & Naumann, C. L. (2009). *Aachener Förderdiagnostische Rechtschreibfehler-Analyse. Systematische Einführung in die Praxis der Fehleranalyse mit Auswertungshilfen zu insgesamt 33 standardisierten Testverfahren als Kopiervorlagen mit Beiträgen von Cordula Löffler* (4., völlig überarbeitete und erweiterte Auflage). Aachen: Alfa Zentaurus. (Siehe S. 19–22, 64, 68, 69).

Hinney, G. (1997). *Neubestimmung von Lehrinhalten für den Rechtschreibunterricht. Ein fachdidaktischer Beitrag zur Schriftaneignung als Problemlöseprozeß*. Frankfurt am Main, Berlin, Bern, New York, Paris, Wien: Peter Lang. (Siehe S. 2, 44, 50, 51, 66).

Hinney, G. (2004). Das Ganze ist mehr als die Summe der Teile. Das Konzept der Schreibsilbe und seine didaktische Modellierung. Ein Beitrag zur Schriftaneignung als Problemlösungsprozess. In U. Bredel, G. Siebert-Ott & T. Thelen (Hrsg.), *Schriftspracherwerb und Orthographie* (Bd. 16, S. 72–90). Diskussionsforum Deutsch. Schneider Verlag Hohengehren. (Siehe S. 50, 51).

Hinney, G. (2010). Wortschreibkompetenz und sprachbewusster Unterricht. Eine Alternativkonzeption zur herkömmlichen Sicht auf den Schriftspracherwerb. In U. Bredel, A. Müller & G. Hinney (Hrsg.), *Schriftsystem und Schrifterwerb: linguistisch – didaktisch – empirisch* (S. 47–100). Berlin, New York: De Gruyter. (Siehe S. 51, 60).

Hinney, G. (2011). Was ist Rechtschreibkompetenz? In U. Bredel & T. Reißig (Hrsg.), *Weiterführender Orthographieerwerb* (Bd. 5, S. 191–225). Deutschunterricht in Theorie und Praxis (DTP). Baltmannsweiler: Schneider Hohengehren. (Siehe S. 50).

Hinney, G. & Menzel, W. (1998). Didaktik des Rechtschreibens. In G. Lange, K. Neumann & W. Ziesenis (Hrsg.), *Taschenbuch des Deutschunterrichts. Grundfragen und Praxis der Sprach- und Literaturdidaktik* (6., vollständig überarbeitete Auflage, Bd. 1, S. 258–304). Schneider Verlag Hohengehren. (Siehe S. 50, 60).

Hinney, G. & Pagel, B. (2007). Rechtschreibkompetenz und Sprachbewusstheit. Ein Unterrichtsprojekt zum forschenden Lernen. *Grundschulunterricht*, (9), 12–24. (Siehe S. 55, 66).

Hornberg, S., Bos, W., Buddeberg, I., Potthoff, B. & Stubbe, T. C. (2007). Anlage und Durchführung von IGLU 2006. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, . . . R. Valtin (Hrsg.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 21–46). Münster, New York, München, Berlin: Waxmann. (Siehe S. 147).

Hornberg, S., Valtin, R., Potthoff, B., Schwippert, K. & Schulz-Zander, R. (2007). Lesekompetenzen von Mädchen und Jungen im internationalen Vergleich. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, . . . R. Valtin (Hrsg.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 195–224). Münster, New York, München, Berlin: Waxmann. (Siehe S. 150, 155).

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der Pädagogischen Diagnostik* (6., neu ausgestattete Auflage). Weinheim, Basel: Beltz. (Siehe S. 80, 81).

Jarsinski, S. (2014). *Quantitative Datenanalyse zur längsschnittlichen Erfassung der Rechtschreibkompetenz in NEPS unter besonderer Berücksichtigung der Kompetenzstruktur und der Einflussfaktoren* (Diss., TU Dortmund). Zugriff unter http://hdl.handle.net/2003/33800. (Siehe S. 55)

Kelava, A. & Moosbrugger, H. (2008). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 73–98). Mit 79 Abbildungen und 43 Tabellen. Heidelberg: Springer. (Siehe S. 97, 98).

Kemmler, L. (1970). *Erfolg und Versagen in der Grundschule. Empirische Untersuchungen* (2. Auflage). Göttingen: Hogrefe. (Siehe S. 2).

Klieme, E. (2004). Was sind Kompetenz und wie lassen sie sich messen? *Zeitschrift für Pädogogik*, *56*(6), 10–13.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., … Vollmer, H. J. (2007). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. Bildungsforschung. Bonn, Berlin: Bundesministerium für Bildung und Forschung. (Siehe S. 119).

Klieme, E. & Hartig, J. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. *Zeitschrift für Erziehungswissenschaft*, *10*(Sonderheft 8), 11–32.

Klieme, E., Hartig, J. & Rauch, D. (2008). The Concept of Competence in Educational Contexts. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of Competencies in Educational Contexts* (S. 3–22). New York: Hogrefe & Huber.

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. *Zeitschrift Pädagogik*, *52*(6), 876–903.

Klieme, E., Merki, K. M. & Hartig, J. (2007). Kompetenzbegriff und Bedeutung von Kompetenzen im Bildungswesen. In J. Hartig & E. Klieme (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung* (Bd. 20, S. 5–15). Bildungsforschung. Bonn, Berlin: Bundesministerium für Bildung und Forschung.

Kluge, F. & Seebold, E. (2002). *KLUGE. Etymologisches Wörterbuch der deutschen Sprache* (24., durchgesehene und erweiterte Auflage). Berlin, New York: de Gruyter.

Koeppen, K., Hartig, J., Klieme, E. & Leutner, D. (2008). Current Issues in Competence Modeling and Assessment. *Journal of Psychology*, *216*(2), 61–73.

Kohn, W. & Öztürk, R. (2013). *Statistik für ökonomen: Datenanalyse mit R und SPSS* (2. Aufl.). Berlin, Heidelberg: Springer. (Siehe S. 113).

Köller, O. (2011). Standardsetzung im Bildungssystem. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Empirische Bildungsforschung. Strukturen und Methoden* (S. 179–192). Wiesbaden: VS Verlag.

Kowalski, K. (2007). *Ergebnisse eines Rechtschreibtests mit Viertklässlern im Rahmen der Voruntersuchung zu IGLU 2006. Theoretische Konzeption und empirische Analyse*. Unveröffentliche Diplomarbeit an der Universität Dortmund. (Siehe S. 4, 56).

Kowalski, K. & Voss, A. (2009). Die IGLU-Ergänzungsstudie 2006 zur Rechtschreibkompetenz von Viertklässlern. In R. Valtin & B. Hofmann (Hrsg.), *Kompetenzmodelle der Orthographie. Empirische Befunde und förderdiagnostische Möglichkeiten* (Bd. 10, S. 26–38). dgLs Beiträge. Berlin: Deutsche Gesellschaft für Lesen und Schreiben. (Siehe S. 72, 112).

Kowalski, K., Voss, A., Valtin, R. & Bos, W. (2010). Erhebungen zur Orthographie in 2001 und 2006: Haben sich die Rechtschreibleistungen verbessert? In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, ... R. Valtin (Hrsg.), *IGLU 2006 -– die Grundschule auf dem Prüfstand. Vertiefende Analysen zu Rahmenbedingungen schulischen Lernens* (S. 33–42). Münster, New York, München, Berlin: Waxmann. (Siehe S. 2, 113, 149).

Krauth, J. (1983). Latente Strukturanalyse. In J. Bredenkamp & H. Feger (Hrsg.), *Strukturierung und Reduzierung von Daten. Forschungsmethoden der Psychologie* (Bd. 4, S. 351–389). Enzyklopädie der Psychologie. Göttingen, Toronto, Zürich: Hogrefe. (Siehe S. 100).

Kultusministerkonferenz, Universität Duisburg-Essen & Institut zur Qualitätsentwicklung im Bildungswesen. (2013). *Kompetenzstufenmodell zu den Bildungsstandards für das Fach Deutsch im Kompetenzbereich „Schreiben", Teilbereich „Rechtschreibung" – Primarbereich –*. Auf Grundlage des Ländervergleichs 2011 überarbeiteter Entwurf in der Version vom 13. Februar 2013. Zugriff am 16.06.2014 unter www.iqb.hu-berlin.de/bista/ksm/KSM_GS_Deutsch_R.pdf. (Siehe S. 28).

Langeheine, R., Pannekoek, J. & van de Pol, F. (1996). Bootstrapping Goodness-of-Fit Measures in Categorical Data Analysis. *Sociological Methods Research*, *24*(4), 492–516. (Siehe S. 106).

Lazarsfeld, P. F. & Henry, N. W. (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin. (Siehe S. 100, 102).

Lehmann, R. H. & Lenkeit, J. (2008). *ELEMENT: Erhebung zum Lese- und Mathematikverständnis – Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin. Abschlussbericht über die Untersuchungen 2003, 2004 und 2005 an Berliner Grundschulen und grundständigen Gymnasien*. Senatsverwaltung für Bildung, Jugend und Sport. Zugriff am 31.07.2013 unter http://www.berlin.de/imperia/md/content/sen-bildung/schulqualitaet/element6_bericht_komplett.pdf. Berlin. (Siehe S. 42).

Lehmann, R. H. & Nikolova, R. (2003). *ELEMENT: Erhebung zum Lese- und Mathematikverständnis – Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin*. Senatsverwaltung für Bildung, Jugend und Sport. Zugriff am 31.07.2013 unter http://www.berlin.de/imperia/md/content/sen-bildung/schulqualitaet/schulleistungsuntersuchungen/element_projektbeschreibung.pdf. Berlin.

Lehmann, R. H. & Nikolova, R. (2005). *ELEMENT: Erhebung zum Lese- und Mathematikverständnis – Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin. Bericht über die Untersuchung 2003 an Berliner Grundschulen und grundständigen Gymnasien*. Senatsverwaltung für Bildung, Jugend und Sport. Zugriff am 31.07.2013 unter http://www.berlin.de/imperia/md/content/sen-bildung/schulqualitaet/schulleistungsuntersuchungen/element_untersuchungsbericht_2003.pdf. Berlin. (Siehe S. 41, 42).

Lehmann, R. H., Peek, R. & Gänsfuß, R. (2011). LAU 5. Aspekte der Lernausgangslage und der Lernentwicklung – Klassenstufe 5 –. Ergebnisse einer längsschnittlichen Untersuchung in Hamburg im September 1996. In B. für Schule und Berufsbildung (Hrsg.), *LAU -– Aspekte der Lernausgangslage und der Lernentwicklung. Klassen-*

*stufen 5, 7 und 9* (Bd. 8, S. 9–120). Hanse – Hamburger Schriften zur Qualität im Bildungswesen. Münster, New York, München, Berlin: Waxmann. (Siehe S. 15).

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Beltz. (Siehe S. 80, 81).

Linacre, J. M. (2002). What do Infit and Outfit Mean-Square and Standardized mean? *Rasch Measurement Transactions*, *16*(2), 878. (Siehe S. 98–100).

Lischeid, T. (2006). Lehren und Lernen mit „LEO". Das Bochumer Projekt der empirischen Lernstands-Ermittlung und –Förderung schulischer Orthografiekompetenz in der Orientierungsstufe weiterführender Schulen (Zwischenbericht Januar 2006). In S. Weinhold (Hrsg.), *Schriftspracherwerb empirisch. Konzepte – Diagnostik – Entwicklung* (Bd. 23, S. 218–233). Diskussionsforum Deutsch. Baltmannsweiler: Schneider Hohengehren. (Siehe S. 36, 37, 65).

Lischeid, T. (2007). Kompetenzorientierung als Prinzip der Einschätzung und Förderung orthografischer Leistungen – das Projekt LEO. In B. Hofmann & R. Valtin (Hrsg.), *Förderdiagnostik im Schriftspracherwerb* (Bd. 6, S. 162–177). dgLs Beiträge. Berlin: Deutsche Gesellschaft für Lesen und Schreiben. (Siehe S. 37).

Lo, Y., Mendell, N. R. & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767–778. (Siehe S. 106).

Löffler, I. & Meyer-Schepers, U. (1992). *Dortmunder Rechtschreibfehler-Analyse zur Ermittlung des Schriftsprachstatus rechtschreibschwacher Schüler*. Dortmund: ILT-Verlag. (Siehe S. 30).

Löffler, I. & Meyer-Schepers, U. (2005). Orthographische Kompetenzen: Ergebnisse qualitativer Fehleranalysen, insbesondere bei schwachen Rechtschreibern. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin & G. Walther (Hrsg.), *IGLU. Vertiefende Analysen zum Leseverständnis, Rahmenbedingungen und Zusatzstudien* (S. 81–108). Münster, New York, München, Berlin: Waxmann. (Siehe S. 31–33, 35–37, 39, 40, 58, 62, 63, 65, 66).

Löffler, I. & Meyer-Schepers, U. (2006a). Dortmunder Schriftkompetenz-Ermittlung (DoSE) – eine modellbasierte Kompetenzdiagnostik für den domänenspezifischen Kompetenzerwerb Rechtschreibung. In W. Bos & S. Hornberg (Hrsg.), *Dokumentation der Fachtagung zu Rechtschreibtests IGLU 2006* (S. 5–10). Dortmund: Internes Papier am Institut für Schulentwicklungsforschung.

Löffler, I. & Meyer-Schepers, U. (2006b). Probleme beim Erwerb von Rechtschreibkompetenz: Ergebnisse qualitativer Fehleranalysen aus IGLU-E. In S. Weinhold (Hrsg.), *Schriftspracherwerb empirisch. Konzepte – Diagnostik – Entwicklung.* (Bd. 23, S. 199–217). Diskussionsforum Deutsch. Baltmannsweiler: Schneider Hohengehren.

Löffler, I. & Meyer-Schepers, U. (2007). Beschreibung von Rechtschreibschwächen mit einem theoretisch fundierten Kompetenzmodell. In B. Hofmann & R. Valtin (Hrsg.), *Förderdiagnostik im Schriftspracherwerb* (Bd. 6, S. 179–196). dgLs Beiträge. Berlin: Deutsche Gesellschaft für Lesen und Schreiben. (Siehe S. 31, 33, 35, 36, 42, 60).

Löffler, I. & Meyer-Schepers, U. (2008). Analyse von Rechtschreibfehlern. Diagnose mit Hilfe von Kompetenzmodellen. *Deutsch differenziert. Fachzeitschrift für die Grundschule, Lese-Rechtschreibschwierigkeiten im weiterführenden Unterricht*, (3), 30–33. (Siehe S. 30–33, 35).

Löffler, I. & Meyer-Schepers, U. (2009). Auswertung nach dem linguistischen Kompetenzmodell. In R. Valtin & B. Hofmann (Hrsg.), *Kompetenzmodelle der Orthographie. Empirische Befunde und förderdiagnostische Möglichkeiten* (Bd. 10, S. 60–73). dgLs Beiträge. Berlin: Deutsche Gesellschaft für Lesen und Schreiben. (Siehe S. 31, 33, 35–37, 73).

Löffler, I., Meyer-Schepers, U. & Schmidt, H. (1990). Sprachwissenschaftlich orientierte Fehleranalyse zur Diagnose einer Lese-Rechtschreibschwäche (Legasthenie). Möglichkeiten zur Planung einer schulischen und außerschulischen Förderung. *Diskussion Deutsch*, (111), 4–17.

Löffler, I., Meyer-Schepers, U. & Stubbe, T. C. (2008). Orthografiekompetenzen von Viertklässlerinnen und Viertklässlern in der Deutschsprachigen Gemeinschaft. In W. Bos, S. Sereni & T. C. Stubbe (Hrsg.), *IGLU Belgien. Lese- und Orthografiekompetenzen von Grundschulkindern in der Deutschsprachigen Gemeinschaft* (S. 137–146). Münster, New York, München, Berlin: Waxmann. (Siehe S. 41).

Lubke, G. H. & Muthén, B. (2005). Investigating Population Heterogeneity With Factor Mixture Models. *Psychological Methods*, *10*(1), 21–39. Zugriff am 30.08.2013 unter https://www.statmodel.com/download/psymeth.pdf. (Siehe S. 100).

Ludlow, L. H. & Haley, S. M. (1995). Rasch Model Logits: Interpretation, Use, and Transformation. *Educational and Psychological Measurement*, *55*(6), 967–975.

Maas, U. (1989). *Grundzüge der deutschen Orthographie* (2. korrigierte und erweiterte Auflage). Osnabrück: Universität Osnabrück, FB Sprach- und Literaturwissenschaft. (Siehe S. 49).

Maas, U. (2006). *Phonologie. Einführung in die funktionale Phonetik des Deutschen* (2., überarbeitete Auflage. Mit zahlreichen Abbildungen und Schautafeln). Studienbucher zur Linguistik. Göttingen: Vandenhoeck & Ruprecht. (Siehe S. 44).

Magidson, J. & Vermunt, J. K. (2004). Latent Class Models. In D. Kaplan (Hrsg.), *The Sage Handbook of Quantitative Methodology for the Social Sciences* (Kap. 10, S. 175–198). Thousand Oaks: Sage. (Siehe S. 100, 106, 163).

Magidson, J. & Vermunt, J. K. (2005). A Nontechnical Introduction to Latent Class Models. Zugriff am 13.09.2013 unter http://statisticalinnovations.com/technicalsupport/lcmodels2.pdf.

Martin, M. O., Kennedy, A. M. & Trong, K. L. (2007). Item Analysis and Review. In M. O. Martin, I. V. S. Mullis & A. M. Kennedy (Hrsg.), *PIRLS 2006 Technical Report* (S. 131–147). Zugriff am 01.08.2013 unter http://timssandpirls.bc.edu/PDF/p06_technical_report.pdf. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College. (Siehe S. 97, 98).

Marx, H. (1997). Literaturüberblick. In F. E. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S. 85–112). Weinheim: Beltz.

Marx, P. (2007). *Lese- und Rechtschreiberwerb*. StandardWissen Lehramt. Paderborn, München, Wien, Zürich: Schöningh. (Siehe S. 2).

Masters, G. N. (1988). The Analysis of Partial Credit Scoring. *Applied Measurement in Education*, *1*(4), 279–297.

May, P. (2000). *Lernförderlichkeit im schriftsprachlichen Unterricht. Effekte des Klassen- und Förderunterrichts in der Grundschule auf den Lernerfolg Ergebnisse der Eva-*

*luation des Projekts „Lesen und Schreiben für alle“ (PLUS)*. Zugriff am 11.02.2008 unter http://www1.uni-hamburg.de/psycholo/frames/projekte/PLUS/PLUS_doc/May01b_PLUS_Evaluation.pdf. Behörde für Schule, Jugend und Berufsbildung. (Siehe S. 16).

May, P. (2001a). *Projekt „Lesen und Schreiben für alle“ (PLUS) Kurzfassung der Ergebnisse der Evaluation*. Zugriff am 11.02.2008 unter http://www1.uni-hamburg.de/psycholo/frames/projekte/PLUS/PLUS_doc/May01a_PLUS_Kurzbericht.pdf. Behörde für Schule, Jugend und Berufsbildung. (Siehe S. 16).

May, P. (2001b). *Prozessbegleitende Evaluation: Lesen und Schreiben in der Grundschule*. Zugriff am 11.02.2008 unter http://www1.uni-hamburg.de/psycholo/frames/projekte/PLUS/May01f.pdf. Behörde für Schule, Jugend und Berufsbildung.

May, P. (2002a). *HSP 1-9. Diagnose orthographischer Kompetenz. Zur Erfassung der grundlegenden Rechtschreibstrategien mit der Hamburger Schreibprobe. Neustandardisierung 2001* (6., aktualisierte und erweiterte Auflage). Manual. Hamburg: vpm. (Siehe S. 2, 12–14, 67, 68).

May, P. (2002b). Lernstandsdiagnose im Rechtschreiben – mit und ohne Test. *Grundschule*, *34*(5), 44–46.

May, P. (2006a). Deutsche Schreibprobe – Webbasierte Erweiterung des Konzepts der Hamburger Schreibprobe. In W. Bos & S. Hornberg (Hrsg.), *Dokumentation der Fachtagung zu Rechtschreibtests IGLU 2006* (S. 11–18). Dortmund: Internes Papier am Institut für Schulentwicklungsforschung. (Siehe S. 12).

May, P. (2006b). Orthographische Kompetenz und ihre Bedingungen am Ende der vierten Jahrgangsstufe. In W. Bos & M. Pietsch (Hrsg.), *KESS 4 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen* (Bd. 1, S. 111–142). Hanse – Hamburger Schriften zur Qualität im Bildungswesen. Münster, New York, München, Berlin: Waxmann. (Siehe S. 11, 15, 16, 149, 150).

May, P. (2008a). Diagnose der orthografischen Kompetenz — von der HSP zur DSP. In W. Schneider, H. Marx & M. Hasselhorn (Hrsg.), *Diagnostik von Rechtschreibleistungen und -kompetenz* (Bd. 6, S. 93–128). Jahrbuch der pädagogisch-psychologischen Diagnostik. Tests und Trends. Göttingen, Bern, Wien, Paris, Oxford, Prag, Toronto, Cambridge, Amsterdam, Kopenhagen: Hogrefe. (Siehe S. 12–14, 17).

May, P. (2008b). schreib.on. Rechtschreibtest im Internet. Ein computergestütztes System zur Diagnose der orthographischen Kompetenz von Kindern, Jugendlichen und Erwachsenen. *Wortspiegel. Fachzeitschrift der LOS*, 10–19. Zugriff am 07.11.2010 unter http://www.dideon.de/sd_dideon.pdf.

May, P. (2009a). Auswertung nach dem Strategiediagnosekonzept. In R. Valtin & B. Hofmann (Hrsg.), *Kompetenzmodelle der Orthographie. Empirische Befunde und förderdiagnostische Möglichkeiten* (Bd. 10, S. 75–89). dgLs Beiträge. Berlin: Deutsche Gesellschaft für Lesen und Schreiben.

May, P. (2009b). Kompetenzen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 6. In W. Bos, M. Bonsen & C. Gröhlich (Hrsg.), *KESS 7 -– Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der*

*Jahrgangsstufe 7* (Bd. 5, S. 59–80). Hanse -– Hamburger Schriften zur Qualität im Bildungswesen. Münster, New York, München, Berlin: Waxmann.

May, P. (2010). Ohografische Kompetenz. In W. Bos & C. Gröhlich (Hrsg.), *KESS 8 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 8* (Bd. 6, S. 67–78). Hanse -– Hamburger Schriften zur Qualität im Bildungswesen. Münster, New York, München, Berlin: Waxmann.

May, P. & Malitzky, V. (1999). Erfassung der Rechtschreibkompetenz in der Sekundarstufe mit der Hamburger Schreibprobe (HSP 4/5 und HSP 5-9). Zugriff am 12.09.2014 unter http://www.peter-may.de/Dokumente/May_doc/MayMal99.pdf. (Siehe S. 72).

May, P., Malitzky, V. & Vieluf, U. (2001). Rechtschreibtests im Vergleich: Wie stellt man deren Güte fest und wie besser nicht? Anmerkungen zur Kritik von Tacke, Völker und Lohmüller an der HSP. *Psychologie in Erziehung und Unterricht*, *48*(2), 146–152.

McCutcheon, A. L. (1987). *Latent Class Analysis*. Newbury Park, Beverly Hills, London, New Delhi: Sage. (Siehe S. 100–102, 105, 106).

McLachlan, G. & Peel, D. (2000). *Finite Mixture Models*. Wiley series in probability and statistics. Applied probability and statistics section. New York, Chichester, Weinheim, Brisbane, Singapore, Toronto: Wiley. (Siehe S. 106).

Merki, K. M. (2009). Kompetenz. In S. Andresen, R. Casale, T. Gabriel, R. Horlacher, S. L. Klee & J. Oelkers (Hrsg.), *Handwörterbuch Erziehungswissenschaft* (S. 492–506). Weinheim, Basel: Beltz.

Meyer-Schepers, U. (1991). *Linguistik und Problematik des Schriftspracherwerbs. Von der Sachlogik des Zusammenhangs von Laut- und Schriftsprache über die Logik der Aneignung von Schriftsprachkompetenz zur Diagnose und Therapie von Fehlersyndromen*. Theorie und Vermittlung der Sprache. Frankfurt am Main, Bern, New York, Paris: Peter Lang. (Siehe S. 30, 31).

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. (2008). Schulgesetz für das Land Nordrhein-Westfalen. Zugriff am 01.07.2009 unter http://www.schulministerium.nrw.de/BP/Schulrecht/Gesetze/SchulG_Info/Schulgesetz.pdf. (Siehe S. 151).

Molenaar, I. W. (1995). Some Background for Item response Theory and the Rasch Model. In G. H. Fischer & I. W. Molenaar (Hrsg.), *Rasch Models: Foundations, Recent Developments, and Applications* (S. 3–14). New York: Springer. (Siehe S. 79, 80, 88).

Moosbrugger, H. (1983). Modelle zur Beschreibung statistischer Zusammenhänge in der psychologischen Forschung. In J. Bredenkamp & H. Feger (Hrsg.), *Strukturierung und Reduzierung von Daten. Forschungsmethoden der Psychologie* (Bd. 4, S. 1–58). Enzyklopädie der Psychologie. Göttingen, Toronto, Zürich: Hogrefe.

Moosbrugger, H. (2008a). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 215–239). Mit 79 Abbildungen und 43 Tabellen. Heidelberg: Springer. (Siehe S. 81, 85).

Moosbrugger, H. (2008b). Klassische Testtheorie (KTT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 99–112). Mit 79 Abbildungen und 43 Tabellen. Heidelberg: Springer. (Siehe S. 80).

Müller, H. (1999). *Probabilistische Testmodelle für diskrete und kontinuierliche Ratingskalen. Einführung in die Item-Response-Theorie für abgestufte und kontinuierliche Items*. Methoden der Psychologie. Mit 15 Abbildungen, 4 Tabellen und Hinweisen auf Software. Bern, Göttingen, Toronto, Seattle: Hans Huber. (Siehe S. 117).

Müller, H. (2000). Summenscore und Trennschärfe beim Rasch-Modell. *Psychologische Rundschau*, *51*(1), 34–35.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M. & Foy, P. (2007). *PIRLS 2006 International Report. IEA's Progress in International Reading Literacy Study in Primary Schools in 40 Countries*. Zugriff am 01.08.2013 unter http://timssandpirls.bc.edu/PDF/PIRLS2006_international_report.pdf. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Munske, H. H. (2005). *Lob der Rechtschreibung. Warum wir schreiben, wie wir schreiben*. München: Verlag C. H. Beck. (Siehe S. 2, 62).

Muthén, B. (2001a). Latent Variable Mixture Modeling. In G. A. Marcoulides & R. E. Schumacker (Hrsg.), *New Developments and Techniques in Structural Equation Modeling* (S. 1–33). Mahwah: Lawrence Erlbaum. (Siehe S. 105).

Muthén, B. (2001b). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class–latent growth modeling. In L. M. Collins & A. G. Sayer (Hrsg.), *New Methods for the Analysis of Change. Decade of Behavior* (S. 291–322). Washington: American Psychological Association.

Muthén, B. & Shedden, K. (1999). Finite Mixture Modeling with Mixture Outcomes Using the EM Algorithm. *Biometrics*, *55*(4), 463–469. Zugriff am 03.09.2013 unter http://www.statmodel.com/bmuthen/articles/Article_078.pdf.

Muthén, L. K. & Muthén, B. O. (2012). *Mplus User's Guide. Statistical Analysis With Latent Variables* (7. Aufl.). Los Angeles. (Siehe S. 100, 105, 106).

Naumann, C. L. (1999). *Orientierungswortschatz. Die wichtigsten Wörter und Regeln für die Rechtschreibung Klasse 1 bis 6*. Weinheim, Basel: Beltz. (Siehe S. 22).

Naumann, C. L. (2008). Zur Rechtschreibkompetenz und ihrer Entwicklung. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Lernstandsbestimmungen im Fach Deutsch. Gute Aufgaben für den Unterricht* (S. 135–161). Weinheim, Basel: Beltz.

Naumann, C. L. & Herné, K.-L. (2008). Vier didaktische Bemerkungen zu Peter Eisenberg und Nanna Fuhrhop in ZS 26 (2007), 15–41: Schulorthographie und Graphematik. *Zeitschrift für Sprachwissenschaft*, *27*(2), 267–271.

Naumann, C. L. & Weinhold, S. (2011). Rechtschreiben. In A. Bremerich-Vos, D. Granzer, U. Behrens & O. Köller (Hrsg.), *Bildungsstandards für die Grundschule: Deutsch konkret. Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen* (3. Aufl., S. 185–201). Cornelsen. (Siehe S. 64, 65).

Nussbeck, F. W., Eid, M. & Geiser, C. (2010). Mischverteilungsmodelle. In H. Holling & B. Schmitz (Hrsg.), *Handbuch Statistik, Methoden und Evaluation* (Bd. 13, S. 562–568). Handbuch der Psychologie. Göttingen, Bern, Wien, Paris, Oxford, Prag, Toronto, Cambridge, Amsterdam, Kopenhagen, Stockholm: Hogrefe. (Siehe S. 106).

Nylund, K. L., Asparouhov, T. & Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation

Study. *Structural Equation Modeling*, *14*(4), 535–569. Zugriff am 22.08.2013 unter http://www.statmodel.com/download/LCA_tech11_nylund_v83.pdf. (Siehe S. 106, 107, 163).

OECD. (2005). *PISA 2003 Technical Report*. Zugriff am 13.07.2011 unter http://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/35188570.pdf. Organisation for Economic Co-operation and Development. (Siehe S. 97, 98, 123).

OECD. (2009). *PISA 2006 Technical Report*. Zugriff am 11.07.2013 unter http://www.oecd.org/pisa/pisaproducts/42025182.pdf. Organisation for Economic Co-operation and Development. (Siehe S. 123).

OECD. (2011). *PISA 2009 Ergebnisse: Was macht eine Schule erfolgreich? Lernumfeld und schulische Organisation in PISA*. Organisation for Economic Co-operation and Development. (Siehe S. 107).

OECD. (2012). *PISA 2009 Technical Report*. Zugriff am 11.07.2013 unter http://www.oecd.org/pisa/pisaproducts/50036771.pdf. Organisation for Economic Co-operation and Development. (Siehe S. 16, 100, 122, 123).

OECD. (2014). *PISA 2012 Technical Report*. Zugriff am 14.05.2015 unter http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf. Organisation for Economic Co-operation and Development. (Siehe S. 16, 123).

Osteen, P. (2010). An Introduction to Using Multidimensional Item Response Theory to Assess Latent Factor Structures. *Journal of the Society for Social Work and Research*, *1*(2), 66–82. (Siehe S. 93, 94, 96, 100, 119).

Pant, H. A., Böhme, K. & Köller, O. (2012). Das Kompetenzkonzept der Bildungsstandards und die Entwicklung von Kompetenzstufenmodellen. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 49–55). Münster, New York, München, Berlin: Waxmann. (Siehe S. 25).

Pant, H. A., Tiffin-Richards, S. P. & Köller, O. (2010). Standard-Setting für Kompetenztests im Large-Scale-Assessment. *Zeitschrift für Pädagogik*, *56*, 175–188.

Pietsch, M. (2007). Soziale Herkunft und Schulleistung Hamburger Kinder am Ende der Grundschulzeit. In W. Bos, C. Gröhlich & M. Pietsch (Hrsg.), *KESS 4 -– Lehr- und Lernbedingungen in Hamburger Grundschulen* (Bd. 2, S. 7–34). Hanse – Hamburger Schriften zur Qualität im Bildungswesen. Münster, New York, München, Berlin: Waxmann. (Siehe S. 155).

Rabe, T. (2012). Vorwort des Präsidenten der Kultusministerkonferenz. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 9–10). Münster, New York, München, Berlin: Waxmann. (Siehe S. 24).

Rasch, B., Friese, M., Hofmann, W. & Naumann, E. (2004). *Quantitative Methoden*. Berlin, Heidelberg, New York, Hongkong, London, Mailand, Paris, Tokio: Springer. (Siehe S. 98, 124, 146).

Rasch, B., Friese, M., Hofmann, W. & Naumann, E. (2006). *Quantitative Methoden* (2., erweiterte Auflage. Mit 29 Abbildungen und 61 Tabellen). Berlin, Heidelberg, New York, Hongkong, London, Mailand, Paris, Tokio: Springer.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Studies in Mathematical Psychology. Copenhagen: Nielsen & Lydiche. (Siehe S. 5, 82, 84, 87, 88).

Rat für deutsche Rechtschreibung. (2006). *Regeln und Wörterverzeichnis. Entsprechend den Empfehlungen des Rats für deutsche Rechtschreibung, überarbeitete Fassung des amtlichen Regelwerks 2004 mit den Nachträgen aus dem Bericht 2010*. Zugriff am 18.01.2015 unter http://rechtschreibrat.ids-mannheim.de/download/regeln2006.pdf. München, Mannheim. (Siehe S. 62).

Rauch, D. & Hartig, J. (2008). Interpretation von Testwerten in der IRT. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 240–250). Mit 79 Abbildungen und 43 Tabellen. Heidelberg: Springer. (Siehe S. 83, 84, 119, 145).

Richter, D., Engelbert, M., Böhme, K., Haag, N., Hannighofer, J., Reimers, H., … Stanat, P. (2012). Anlage und Durchführung des Ländervergleichs. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 85–102). Münster, New York, München, Berlin: Waxmann. (Siehe S. 25).

Röber-Siekmeyer, C. (1999). *Ein anderer Weg zur Groß- und Kleinschreibung*. Leipzig, Stuttgart, Düsseldorf: Klett-Grundschulverlag. (Siehe S. 49).

Rost, J. (2000). Haben ordinale Rasch-Modelle variierende Trennschärfen? Eine Antwort auf die Wiener Repliken. *Psychologische Rundschau*, *51*(1), 36–37.

Rost, J. (2001). The Growing Family of Rasch Models. In A. Boomsma, M. A. J. van Duijn & T. A. B. Snijders (Hrsg.), *Essays on Item Response Theory* (Bd. 157, S. 25–42). Lecture Notes in Statistics. New York, Berlin, Heidelberg, Barcelona, Hong Kong, London, Milan, Paris, Singapore, Tokyo: Springer. (Siehe S. 88).

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2., vollständig überarbeitete und erweiterte Auflage). Bern, Göttingen, Toronto, Seattle: Verlag Hans Huber. (Siehe S. 5, 76, 78–86, 88, 91–96, 98, 101–106, 117, 122, 124, 127, 138, 145, 163).

Rost, J. (2006). Latent-Class-Analyse. In F. Petermann & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (Bd. 4, S. 275–287). Handbuch der Psychologie. Göttingen, Bern, Wien, Toronto, Seattle, Oxford, Prag: Hogrefe. (Siehe S. 102, 103, 107).

Rost, J. & Langeheine, R. (1997). A Guide through Latent Structure Models for Categorical Data. In *Applications of Latent Trait and Latent Class Models in the Social Sciences* (Kap. 1, S. 13–37). Münster, New York, München, Berlin: Waxmann. (Siehe S. 101, 104, 105).

Rost, J., Walter, O., Carstensen, C. H., Senkbeil, M. & Prenzel, M. (2004). Naturwissenschaftliche Kompetenz. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, … U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland -– Ergebnisse des zweiten internationalen Vergleichs* (S. 111–146). Münster, New York, München, Berlin: Waxmann. (Siehe S. 89, 153).

Rutkowski, L., Gonzalez, E., Joncas, M. & von Davier, M. (2010). International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting. *Educational Researcher*, *39*(2), 142–151.

Ryan, J. P. (1983). Introduction to latent trait analysis and item response theory. In *Testing in the Schools. New Directions for Testing and Measurement* (S. 48–65). San Francisco: Jossey-Bass. (Siehe S. 79, 81, 83, 87).

Samuelson, K. & Raczynski, K. (2013). Latent Class/Profile Analysis. In Y. M. Petscher, C. Schatschneider & D. L. Compton (Hrsg.), *Applied Quantitative Analysis in Education and the Social Sciences* (S. 304–328). New York: Routledge. (Siehe S. 106, 164).

Scheele, V. (2006a). *Entwicklung fortgeschrittener Rechtschreibfertigkeiten. Ein Beitrag zum Erwerb der „orthographischen" Strategien*. Theorie und Vermittlung der Sprache. Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang.

Scheele, V. (2006b). Praktische Entwicklung fortgeschrittener Rechtschreibfertigkeiten: Zum Erwerb der ‚orthographischen' Strategien. In S. Weinhold (Hrsg.), *Schriftspracherwerb empirisch. Konzepte – Diagnostik – Entwicklung.* (Bd. 23, S. 234–260). Diskussionsforum Deutsch. Baltmannsweiler: Schneider Hohengehren. (Siehe S. 12).

Schermelleh-Engel, K., Kelava, A. & Moosbrugger, H. (2006). Gütekriterien. In F. Petermann & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (Bd. 4, S. 420–433). Handbuch der Psychologie. Göttingen, Bern, Wien, Toronto, Seattle, Oxford, Prag: Hogrefe. (Siehe S. 139).

Schipolowski, S. & Böhme, K. (2010). Der Ländervergleich im Fach Deutsch. In O. Köller, M. Knigge & B. Tesch (Hrsg.), *Sprachliche Kompetenzen im Ländervergleich* (S. 87–97). Münster, New York, München, Berlin: Waxmann.

Schneider, W. (2008a). Entwicklung der Schriftsprachkompetenz vom frühen Kindes- bis zum frühen Erwachsenenalter. In W. Schneider (Hrsg.), *Entwicklung von der Kindheit bis zum Erwachsenenalter: Befunde der Münchner Längsschnittstudie LOGIK* (S. 167–186). Weinheim: Beltz. (Siehe S. 2).

Schneider, W. (2008b). Entwicklung und Erfassung der Rechtschreibkompetenz im Jugend- und Erwachsenenalter. In W. Schneider, H. Marx & M. Hasselhorn (Hrsg.), *Diagnostik von Rechtschreibleistungen und -kompetenz* (Bd. 6, S. 145–158). Jahrbuch der pädagogisch-psychologischen Diagnostik. Tests und Trends. Göttingen, Bern, Wien, Paris, Oxford, Prag, Toronto, Cambridge, Amsterdam, Kopenhagen: Hogrefe. (Siehe S. 2).

Schneider, W. & Stefanek, J. (2007). Entwicklung der Rechtschreibleistung vom frühen Schul- bis zum frühen Erwachsenenalter. Längsschnittliche Befunde der Münchner LOGIK-Studie. *Zeitschrift für Pädagogische Psychologie*, *21*(1), 77–82. (Siehe S. 1, 2).

Schneider, W., Stefanek, J. & Dotzler, H. (1997). Erwerb des Lesens und des Rechtschreibens. Ergebnisse aus dem SCHOLASTIK-Projekt. In F. E. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S. 113–130). Weinheim: Beltz.

Schönweiss, F. (2006). Schulentwicklung mit dem Lernserver der Uni Münster: Computergestützte Diagnose, Förderung und Fortbildung. In W. Bos & S. Hornberg (Hrsg.), *Dokumentation der Fachtagung zu Rechtschreibtests IGLU 2006* (S. 19–23). Dortmund: Internes Papier am Institut für Schulentwicklungsforschung.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.). (2005a). *Bildungsstandards der Kultusministerkonferenz. Erläuterungen zur Konzeption und Entwicklung*. Beschlüsse der Kultusministerkonferenz. München: Luchterhand.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.). (2005b). *Bildungsstandards im Fach Deutsch für den Primarbereich. Beschluss vom 15.10.2004*. Beschlüsse der Kultusministerkonferenz. München: Luchterhand. (Siehe S. 2).

Stanat, P., Pant, H. A., Richter, D., Böhme, K., Engelbert, M., Haag, N., … Weirich, S. (2012). Der Blick in die Länder. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 131–172). Münster, New York, München, Berlin: Waxmann. (Siehe S. 25).

Steinig, W. & Betzel, D. (2014). Schreiben Grundschüler heute schlechter als vor 40 Jahren? Texte von Viertklässlern aus den Jahren 1972, 2002 und 2012. In A. Plewnia & A. Witt (Hrsg.), *Sprachverfall? Dynamik – Wandel – Variation* (S. 353–371). Jahrbuch des Instituts für Deutsche Sprache 2013. Berlin, Boston: de Gruyter. (Siehe S. 1).

Steinig, W., Betzel, D., Geider, F. J. & Herbold, A. (2009). *Schreiben von Kindern im diachronen Vergleich. Texte von Viertklässlern aus den Jahren 1972 und 2002*. Münster, New York, München, Berlin: Waxmann. (Siehe S. 1).

Strietholt, R., Naujokat, K., Mai, T., Kretschmer, S., Jarsinski, S., Goy, M., … Blatt, I. (2013). The National Educational Panel Study (NEPS) in Germany: an overview of design, research options and access, with a focus on lower-secondary school. *European Educational Research Journal*, *12*(4), 568–579. (Siehe S. 77, 178).

Strobl, C. (2010). *Das Rasch-Modell. Eine verständliche Einführung für Studium und Praxis*. Sozialwissenschaftliche Forschungsmethoden. München, Mering: Rainer Hampp Verlag. (Siehe S. 84, 86–89).

Stubbe, T. C., Bos, W. & Hornberg, S. (2008). Soziale und kulturelle Disparitäten der Schülerleistungen in den Ländern der Bundesrepublik Deutschland. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, …R. Valtin (Hrsg.), *IGLU-E 2006. Die Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich* (S. 103–109). Münster, New York, München, Berlin: Waxmann.

Stubbe, T. C., Sereni, S. & Bos, W. (2008). Anlage und Durchführung der Internationalen Grundschul-Lese-Untersuchung in der Deutschsprachigen Gemeinschaft (IGLU Belgien). In W. Bos, S. Sereni & T. C. Stubbe (Hrsg.), *IGLU Belgien. Lese- und Orthografiekompetenzen von Grundschulkindern in der Deutschsprachigen Gemeinschaft* (S. 19–40). Münster, New York, München, Berlin: Waxmann.

Tarelli, I., Valtin, R., Bos, W., Bremerich-Vos, A. & Schwippert, K. (2012). IGLU 2011: Wichtige Ergebnisse im überblick. In W. Bos, I. Tarelli, A. Bremerich-Vos & K. Schwippert (Hrsg.), *IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 11–26). Münster, New York, München, Berlin: Waxmann. (Siehe S. 107).

Tarelli, I., Wendt, H., Bos, W. & Zylowski, A. (2012). Ziele, Anlage und Durchführung der Internationalen Grundschul-Lese-Untersuchung (IGLU 2011). In W. Bos, I. Tarelli, A. Bremerich-Vos & K. Schwippert (Hrsg.), *IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 27–68). Münster, New York, München, Berlin: Waxmann. (Siehe S. 156).

Thomé, G. (2000). Linguistische und psycholinguistische Grundlagen der Orthografie: Die Schrift und das Schreibenlernen. In R. Valtin (Hrsg.), *Rechtschreiben lernen in den Klassen 1 - 6. Grundlagen und didaktische Hilfen* (S. 12–16). Frankfurt am Main: Grundschulverband – Arbeitskreis Grundschule e.V. (Siehe S. 62, 64).

Thomé, G. (2006). Entwicklung der basalen Rechtschreibkenntnisse. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch* (2., durchgesehene Auflage, Bd. 1, S. 368–379). Paderborn, München, Wien, Zürich: Ferdinand Schöningh. (Siehe S. 64).

Thomé, G. & Eichler, W. (2008). Rechtschreiben Deutsch. In DESI-Konsortium (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (S. 104–111). Weinheim, Basel: Beltz. (Siehe S. 152).

Thomé, G. & Gomolka, J. (2007). Rechtschreiben. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (S. 140–146). Weinheim, Basel: Beltz.

Thomé, G. & Thomé, D. (2000). Sind quantitative Tests und Methoden heute noch zeitgemäß? Probleme der Rechtschreibdiagnostik. In R. Valtin (Hrsg.), *Rechtschreiben lernen in den Klassen 1-6: Grundlagen und didaktische Hilfen* (S. 120–123). Frankfurt am Main: Arbeitskreis Grundschule.

Thomé, G. & Thomé, D. (2004a). Der orthographische Fehler zwischen Orthographietheorie und Entwicklungspsychologie. Aspekte der qualitativen Fehleranalyse und Förderdiagnostik. In A. Bremerich-Vos, C. Löffler & K.-L. Herné (Hrsg.), *Neue Beiträge zur Rechtschreibtheorie und -didaktik* (S. 163–178). Freiburg: Fillibach.

Thomé, G. & Thomé, D. (2004b). Die Oldenburger Fehleranalyse (OLFA). Ein Instrument zur Ermittlung der Rechtschreibkompetenz ab Klasse 3 und zur Qualitätssicherung von Rechtschreibfördermaßnahmen. In G. Thomé (Hrsg.), *Lese-Rechtschreibschwierigkeiten (LRS) und Legasthenie. Eine grundlegende Einführung* (S. 128–142). Weinheim, Basel: Beltz.

Thomé, G. & Thomé, D. (2004c). *Oldenburger Fehleranalyse OLFA: Instrument und Handbuch zur Ermittlung der orthographischen Kompetenz aus freien Texten ab Klasse 3 und zur Qualitätssicherung und Fördermaßnahmen*. Oldenburg: Igel Verlag Wissenschaft.

Thomé, G. & Thomé, D. (2010). *OLFA 3-9: Oldenburger Fehleranalyse für die Klassen 3–9. Instrument und Handbuch zur Ermittlung der orthographischen Kompetenz und Leistung aus freien Texten und für die Planung und Qualitätssicherung von*

*Fördermaßnahmen (mit Kopiervorlagen).* (2., erweiterte und verbesserte Auflage). Oldenburg: Institut für sprachliche Bildung. (Siehe S. 20, 64).

Valtin, R. (1997). Kommentar. In F. E. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S. 131–138). Weinheim: Beltz.

Valtin, R., Badel, I., Löffler, I., Meyer-Schepers, U. & Voss, A. (2003). Orthographische Kompetenzen von Schülerinnen und Schülern der vierten Klasse. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin & G. Walther (Hrsg.), *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich* (S. 227–264). Münster, New York, München, Berlin: Waxmann. (Siehe S. 2, 3, 29–33, 35, 36, 38, 59, 62, 63, 66, 149, 152, 155).

Valtin, R., Bos, W., Buddeberg, I., Goy, M. & Potthoff, B. (2008). Lesekompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe im nationalen und internationalen Vergleich. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, . . . R. Valtin (Hrsg.), *IGLU-E 2006. Die Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich* (S. 51–101). Münster, New York, München, Berlin: Waxmann. (Siehe S. 150).

Valtin, R., Löffler, I., Meyer-Schepers, U. & Badel, I. (2004a). Orthographische Kompetenzen von Schülerinnen und Schülern der vierten Klasse im Vergleich der Länder. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin & G. Walther (Hrsg.), *IGLU. Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich* (S. 141–164). Münster, New York, München, Berlin: Waxmann. (Siehe S. 3, 31–33, 35, 36, 59, 62, 65, 66).

Valtin, R., Löffler, I., Meyer-Schepers, U. & Badel, I. (2004b). Was Sie schon immer über den Rechtschreibunterricht wissen wollten und aus IGLU-E erfahren können. *Grundschulunterricht*, *51*(4), 2–5.

Valtin, R., Meyer-Schepers, U. & Löffler, I. (2003). Rechtschreiben – ein wahres Schulkreuz? *Grundschule*, *35*(12), 40–41.

van Ackeren, I. & Klemm, K. (2011). *Entstehung, Struktur und Steuerung des deutschen Schulsystems. Eine Einführung* (2., aktualisierte und überarbeitete Auflage). Wiesbaden: VS Verlag. (Siehe S. 151).

Vermunt, J. K. & Magidson, J. (2002). Latent Class Cluster Analysis. In J. A. Hagenaars & A. L. McCutcheon (Hrsg.), *Applied Latent Class Analysis* (S. 89–106). Cambridge: Cambridge University Press. (Siehe S. 100, 106).

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).* Empirische Erziehungswissenschaft. Münster, New York, München, Berlin: Waxmann.

Voss, A. (2007). Neue Wege zum Rechtschreiblernen. IGLU 2006 - Voruntersuchung: Rechtschreibkompetenz am Ende der Grundschule. *Praxis Deutsch*, *201*(34), 58–59.

Voss, A. (2009). Zur Erfassung und Modellierung von Rechtschreibkompetenz. In R. Valtin & B. Hofmann (Hrsg.), *Kompetenzmodelle der Orthographie. Empirische Befunde und förderdiagnostische Möglichkeiten* (Bd. 10, S. 12–24). dgLs Beiträge. Berlin: Deutsche Gesellschaft für Lesen und Schreiben.

Voss, A., Blatt, I., Gebauer, M. M., Müller, A. & Masanek, N. (2008). Unterrichtsentwicklung als integrierte Schulentwicklung. Das Hamburger Leseförderprojekt (HeLp). In W. Bos, H. G. Holtappels, H. Pfeiffer, H.-G. Rolff & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung. Daten, Beispiele und Perspektiven* (Bd. 15, S. 93–122). Jahrbuch der Schulentwicklung. Weinheim, München: Juventa.

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. (Siehe S. 4, 51–53, 56, 57, 59, 64, 66, 149, 155).

Voss, A., Löffler, I., Meyer-Schepers, U., Meckel, C. & Kowalski, K. (2008). Frühdiagnose rechtschreibschwächerer Schülerinnen und Schüler auf der Grundlage von Kompetenzmodellen. Die Analyse von Lernentwicklungsverläufen als Aufgabe schulischer Effektivitätsforschung. In W. Bos, H. G. Holtappels, H. Pfeiffer, H.-G. Rolff & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung. Daten, Beispiele und Perspektiven* (Bd. 15, S. 123–156). Jahrbuch der Schulentwicklung. Weinheim, München: Juventa. (Siehe S. 31–33, 35–37, 42, 43, 60, 65, 73, 77).

Walter, O. (2005). *Kompetenzmessung in den PISA-Studien: Simulationen zur Schätzung von Verteilungsparametern und Reliabilitäten*. Lengerich, Berlin, Bremen, Miami, Riga, Viernheim, Wien, Zagreb: Pabst. (Siehe S. 83, 90, 91, 145).

Walter, O. & Rost, J. (2011). Psychometrische Grundlagen von Large Scale Assessments. In L. F. Hornke, M. Amelang & M. Kersting (Hrsg.), *Methoden der Psychologischen Diagnostik. Psychologische Diagnostik* (Bd. 2, S. 87–149). Enzyklopädie der Psychologie. Göttingen, Bern, Toronto, Seattle: Hogrefe. (Siehe S. 89).

Wang, J. & Wang, X. (2012). *Structural equation modeling : applications using Mplus*. Wiley series in probability and statistics. Chichester: Wiley. (Siehe S. 100, 107, 161, 164).

Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Models. *Psychometrika*, *54*(3), 427–450. (Siehe S. 145).

Weinert, F. E. (1999). *Definition and Selection of Competencies. Concepts of Competence*. Max Planck Institute for Psychological Research. Gutachten zum OECD-Projekt „Definition and Selection of Competencies“: Theoretical and Conceptual Foundations (DeSeCo). Munich.

Weinert, F. E. (2001a). Concept of Competence: A Conceptual Clarification. In D. S. Rychen & L. H. Salganik (Hrsg.), *Defining and Selecting Key Competencies* (S. 56–66). Seattle, Toronto, Bern, Göttingen: Hogrefe & Huber.

Weinert, F. E. (2001b). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Weinheim, Basel: Beltz.

Weirich, S., Haag, N. & Roppelt, A. (2012). Testdesign und Auswertung des Ländervergleichs: technische Grundlagen. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 277–290). Münster, New York, München, Berlin: Waxmann. (Siehe S. 25).

Wendt, H., Tarelli, I., Bos, W., Frey, K. & Vennemann, M. (2012). Ziele, Anlage und Durchführung der Trends in International Mathematics and Science Study (TIMSS 2011). In W. Bos, H. Wendt, O. Köller & C. Selter (Hrsg.), *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 27–68). Münster, New York, München, Berlin: Waxmann. (Siehe S. 145).

Wilson, M. (2005). *Constructing Measures. An Item Response Modeling Approach*. Mahwah: Lawrence Erlbaum. (Siehe S. 97, 99, 100).

Wilson, M., Boeck, P. D. & Carstensen, C. H. (2008). Explanatory Item Response Models: A Brief Introduction. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of Competencies in Educational Contexts* (S. 91–120). New York: Hogrefe & Huber.

Winkelmann, H. & Böhme, K. (2009). Anlage und Durchführung der Pilotierung der Bildungsstandards. 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. 31–41). Weinheim, Basel: Beltz. (Siehe S. 98).

Wolff, H.-G. & Bacher, J. (2010). Hauptkomponentenanalyse und explorative Faktorenanalyse. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 333–365). Wiesbaden: VS Verlag. (Siehe S. 24).

Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis. Rasch Measurement*. Chicago: Mesa Press. (Siehe S. 97, 99).

Wright, B. D. & Stone, M. H. (1979). *Best Test Design. Rasch Measurement*. Chicago: Mesa Press. (Siehe S. 83, 97).

Wu, M. & Adams, R. (2006). Modelling Mathematics Problem Solving Item Responses Using a Multidimensional IRT Model. *Mathematics Education Research Journal*, *18*(2), 93–113. Zugriff am 22.02.2013 unter http://www.merga.net.au/documents/MERJ_18_2_Wu.pdf. (Siehe S. 89, 93, 122, 125).

Wu, M. & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Zugriff am 09.05.2011 unter http://www.edmeasurement.com.au/_publications/RaschMeasurement_Complete.pdf. Melbourne: Educational Measurement Solutions. (Siehe S. 84, 91, 96–98, 126, 127, 138).

Wu, M., Adams, R., Wilson, M. & Haldane, S. (2007). *ACER ConQuest. Version 2.0. Generalised Item Response Modelling Software*. Camberwell: Acer Press. (Siehe S. 91, 94, 98, 117, 121, 145).

Yang, C.-C. (2006). Evaluating latent class analysis models in qualitative phenotype identification. *Computational Statistics & Data Analysis*, *50*(4), 1090–1104. (Siehe S. 106).