Bibliographischer Hinweis sowie Verlagsrechte bei den online-Versionen der DD-Beiträge:

Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur
http://www.didaktik-deutsch.de
15. Jahrgang 2010 – ISSN 1431-4355
Schneider Verlag Hohengehren GmbH

Katrin Böhme/Albert Bremerich-Vos

HABEN WIR TOMATEN AUF DEN AUGEN?
Eine Replik auf den Beitrag von Carl Ludwig Naumann & Karl-Ludwig Herné

In: Didaktik Deutsch. Jg. 15. H. 28. S. 22-31.

Katrin Böhme/Albert Bremerich-Vos

Haben wir Tomaten auf den Augen?
Eine Replik auf den Beitrag von Carl Ludwig Naumann & Karl-Ludwig Herné

Im Folgenden gehen wir im Wesentlichen auf vier Aspekte der Argumentation von Naumann und Herné ein: Zunächst beschäftigen wir uns mit der Frage, inwiefern eine große Schulleistungsstudie wie die, auf die sich die beiden Autoren beziehen, in individualdiagnostischer Perspektive interpretiert werden kann und sollte (1). Dann interessieren Art und Anzahl der Kategorien, anhand derer Lupenstellen in den Testwörtern bestimmt werden (2). Davon hängt – drittens – ab, wie die Komplexität der Testwörter bzw. Items zu fassen ist und wie die Hypothese beurteilt werden kann, dass die Wörter umso schwieriger werden, je mehr Lupenstellen sie enthalten. Hinzu kommt die Frage nach der „Interaktion" zwischen Lupenstellen (3). Abschließend thematisieren wir die Frage der Dimensionalität des Konstrukts der Rechtschreibkompetenz (4).

1 Large-Scale Assessment und Individualdiagnostik

Naumann & Herné zitieren aus dem Beitrag von Böhme & Bremerich-Vos (2009) u. a. den Satz, „dass alle hier vorgestellten Ergebnisse […] keine Aussagen über intraindividuelle Leistungsentwicklungen gestatten." (353) Gerade dieser Aspekt sollte nach Ansicht der beiden Autoren jedoch im Mittelpunkt stehen: „Für die Lernenden und die Lehrkräfte muss es aber gerade um »intraindividuelle Lernentwicklungen« gehen. Daher wäre es misslich, wenn Lernstandsbestimmung und Lernwege nicht aufeinander bezogen werden könnten, wie es das Zitat nahelegt." (Naumann & Herné, in diesem Heft, 7) Zunächst möchten wir darauf hinweisen, dass es uns nicht möglich war, die *Entwicklung* der Rechtschreibkompetenz zu betrachten, da es sich nicht um eine längsschnittliche Untersuchung handelte. Es wurden zwar Dritt- und Viertklässler untersucht, aber jeweils nur querschnittlich.

Unstrittig ist, dass standardbezogene Tests wie der unsrige *im Prinzip* auch der Individualdiagnostik dienen können, worauf wir hingewiesen haben (Böhme & Bremerich-Vos, 336). Allein, dies war nicht die Zielstellung der zitierten Untersuchung. Zentral waren die Erprobung von Aufgaben und die Ermittlung von bundesweit gültigen Normwerten, die eine spätere Evaluierung der Bildungsstandards im Sinne eines Systemmonitorings gestatten. Außerdem können – aufgrund von anders gelagerten Zielstellungen – in aktuellen großen Schulleistungsstudien für die Individualebene oftmals keine hinreichend reliablen Messungen der Kompetenzstände *einzelner* Schülerinnen und Schüler erreicht werden. Das hat mit der bewusst angestrebten inhaltlichen Breite der Messung zu tun. In unserer Studie wurde ein Multi-Matrix-Design eingesetzt, sodass jede einzelne Schülerin und jeder einzelne Schüler nur wenige Wörter zu schreiben hatte. In einem solchen Design bearbeiten verschiedene Stu-

dienteilnehmer unterschiedliche Sets von Items, darüber hinaus fungieren einige Items als „Anker“, d. h. sie kommen in unterschiedlichen Testheften und damit in unterschiedlichen Itemsets vor. Auch wenn auf diese Weise mehr Items im Spiel sind: Auf Individualebene ist mit einem erheblichen Messfehler zu rechnen. Er fällt auf dieser Ebene viel stärker ins Gewicht als auf der Ebene von Klassen oder Schulen. Somit ist aufgrund des Studiendesigns hinsichtlich der individualdiagnostischen Reliabilität der Kompetenzfeststellung Vorsicht geboten. Eine Interpretation der Ergebnisse in dem von Naumann & Herné gewünschten Sinne ist kaum zuverlässig möglich. (I)

2 Art und Anzahl von Lupenstellenkategorien

Einigkeit besteht darin, dass eine qualitative Fehleranalyse nottut, dass Fehler nicht (nur) wortbezogen, sondern (auch) auf der Ebene von Lupenstellen zu betrachten sind. In unserer Studie haben wir auf gängige „Stufen“-Theorien des Orthografieerwerbs Bezug genommen. Man kann davon ausgehen, dass die alphabetische „Stufe“ für Kinder der dritten und vierten Klasse in der Regel kaum noch Probleme aufwirft. Angesichts dessen haben wir uns für Fehlerkategorien entschieden, die primär mit der orthografischen „Stufe“ assoziiert sind. Diese Kategorien, die von Naumann & Herné referiert werden (6), sind in Anlehnung an die Aachener Förderdiagnostische Rechtschreibfehler-Analyse (AFRA) (Herné & Naumann 2002) gebildet worden. In AFRA werden 25 Fehlerkategorien unterschieden; darunter sind allein acht Kategorien, die sich auf die Phonem-Graphem-Korrespondenz beziehen (u. a. Buchstaben-Form, Graphem-Auswahl, Graphem-Folge, Fremdwort-Grapheme). Für unsere Zwecke haben wir die Anzahl der Kategorien auf neun reduziert, wobei zum Teil zwei Kategorien zu einer zusammengefasst oder Kategorien modifiziert wurden. Diese Reduktion hatte u. a. testökonomische Gründe: Immerhin waren die Testergebnisse von knapp 3.500 Schülerinnen und Schülern auszuwerten. Sollen im Test im Wesentlichen Aspekte relevant sein, die mit der orthografischen „Stufe“ assoziiert sind, dann kann die Klassifikation der Fehler nicht erschöpfend sein. Es ist also wichtig zu wissen, wie groß der Anteil der Fehler ist, die mit auf diese „Stufe“ bezogenen Kategorien nicht erfasst werden können. Diese Fehler wurden mit einer zehnten Kategorie als „anderer Fehler“ (AF) signiert.Wenn die Häufigkeit eines anderen Fehlers bei einem Wort bei zehn Prozent liegt, so besagt dies, dass ein Zehntel der Schülerinnen und Schüler hier zusätzlich zu den Fehlern an den Lupenstellen an anderen Stellen fehlerhaft geschrieben hat. In diesem Kontext ist ein Befund relevant, auf den Naumann & Herné nicht eingehen: Analysiert man alle 80 Testwörter unserer Studie, dann beträgt die Häufigkeit eines mit „AF“ zu signierenden Fehlers zwischen null und 58 Prozent. Kumulativ betrachtet: Bei 25 von 80 Wörtern wurde für weniger als 10 Prozent der Kinder „AF“ kodiert, bei 53 für weniger als 20 Prozent, bei 67 Wörtern unter 30 Prozent. Nur bei elf der 80 Wörter gab es eine Häufigkeit von „anderen Fehlern“ zwischen 30 und 44 Prozent (Böhme & Bremerich-Vos 2009, 343). Daraus kann man schließen, dass die weit überwiegende Mehrzahl der

Fehler tatsächlich anhand der neun von uns vorgesehenen Fehlerkategorien zu erfassen ist. (II)

Man kann postulieren, dass eine (Fehler-)Klassifikation exhaustiv sein sollte, dass also *alle* qualitativ plausibel unterscheidbaren Fehler kategorial erfasst werden sollten. Dem lässt sich entgegenhalten, dass es u. a. aus pragmatischen Gründen sinnvoll sein mag, den Auflösungsgrad bzw. die „Korngröße" der Klassifikation zu variieren. Naumann & Herné schlagen 13 Kategorien (ohne AF) vor mit der Konsequenz, dass sich die Anzahl der Lupenstellen der Testwörter fast verdoppelt (von 54 auf 102). Angesichts der Befunde, welche die Häufigkeit eines „anderen Fehlers" betreffen, bleibt für uns fraglich, ob der Informationsgewinn, der mit der Verdopplung der Zahl der Fehlerlupenstellen verbunden ist, den massiv erhöhten testökonomischen Aufwand zu rechtfertigen vermag.

Was die Definitionen der Fehlerkategorien angeht, so bleiben u. E. bei *jeder* Klassifikation dezisionistische „Reste", über die im Einzelnen zu diskutieren wäre. Naumann & Herné fassen z. B. – einerseits – *nur* die graphisch markierte Länge des Vokals als Lernproblem auf (15); andererseits halten sie dafür, dass bei *jedem* /ɛ/ bzw. /oi/ die Frage nach der Ableitbarkeit gestellt werden muss (16). Wir dagegen begreifen die Länge als solche als Lernproblem und sehen die Kategorie der vokalischen Ableitung nicht für jedes /ɛ/ bzw. /oi/ vor. Wenn man will, kann man beide Versionen kritisieren: Wieso die jeweilige Diskrepanz, wenn man davon ausgeht, dass als Mehrheits- bzw. „Normalschreibung" jeweils der einfache Vokalbuchstabe bzw. die Schreibung als <e> bzw. <eu> gelehrt wird? Dass bei einem Zweifel im zweiten Fall eine (Ableitungs-)Operation weiterhilft, im ersten nicht, kann diese Ungleichbehandlung nicht rechtfertigen. (III) Ein weiteres Beispiel: Wir haben <ver> als häufiges Morphem verbucht (irrtümlicherweise allerdings nicht in <Verkehr>); Naumann & Herné sehen hierfür die Kategorie „Spezielles Graphem" vor. In der Kurzbeschreibung von AFRA (2002) ist unter dem Titel „Spezielle Grapheme Minderheit" <foll> als Falschschreibung aufgelistet, unter dem Etikett „Unselbstständige Morpheme" <Ferbot>. Insofern ist unsere Version mit dieser Beschreibung kompatibel und wieder kann man argumentieren, dass die Zuordnung von Annahmen über die „Standardlehre" abhängt: Wir gehen davon aus, dass in den Jahrgangsstufen 3 und 4 die Zerlegung von Wörtern in „Bausteine" bereits Thema war und dass die Schülerinnen und Schüler häufiger Gelegenheit hatten, <ver> als Baustein zu identifizieren. (IV)

3 Komplexität von Testwörtern und Interaktion von Lupenstellen

Naumann & Herné kommt es vor allem auf die „Itemkomplexität" und die „Interaktionen zwischen Lupenstellen" an. Als Maß für erstere schlagen die beiden Autoren die Zahl der Lupenstellen in einem Wort vor. Je größer die Zahl der Lupenstellen, umso komplexer das Item. Sie ermitteln für unsere Testwörter, welche maximal (bei „Schlittschuhläufer") fünf Lupenstellen vorsehen, Korrelationen der Lupenstellenränge und der Ränge der Lösungshäufigkeiten von -.06 für Drittklässler und

von -.05 für Viertklässler.[1] Es besteht also kein Zusammenhang. Für die von ihnen vorgeschlagene Fehlerklassifikation mit der doppelten Anzahl an Lupenstellen ergeben sich dagegen jeweils Korrelationen von -.44 für Drittklässler und von -.52 für Viertklässler, die als statistisch bedeutsam gekennzeichnet werden. Dies bedeutet: Je mehr Lupenstellen, umso geringer die Lösungshäufigkeit auf Wortebene.

Dieser Befund ist relevant. Schließlich handelt es sich um 20 bzw. 25 Prozent gemeinsamer Varianz der beiden Variablen „Anzahl der Lupenstellen“ und „Lösungshäufigkeit“.

Wenn man diese Korrelationen interpretiert, sollte man aber bedenken, dass die Lösungshäufigkeit von der Anzahl der Lupenstellen womöglich nicht „direkt“ kausal beeinflusst wird, sondern dass auch andere Variablen als verursachend angenommen werden können, z. B. die Wortlänge, gemessen als Anzahl von Buchstaben oder auch Graphemen. (V) Eine Prüfung der ersten dieser beiden Hypothesen ergab, dass der Zusammenhang zwischen Buchstabenzahl und Schwierigkeit zwar tendenziell in dieselbe Richtung weist (für Klasse 3: r = -.17; für Klasse 4: r = -.21), jedoch nicht statistisch abgesichert werden kann (p = .26 bzw. p = .21). Ähnliche Befunde zeigen sich für den Zusammenhang der Graphemzahl mit der Itemschwierigkeit (für Klasse 3: r = -.21; für Klasse 4: r = -.24). Auch diese Korrelationen verfehlen knapp die statistische Signifikanz (p = .20 vs. p = .17).

Interessant ist der Befund, dass die von Naumann & Herné für ihr Kategoriensystem ermittelten Korrelationen sensitiv für die Schwierigkeit der Items sind. Berücksichtigt man nämlich die drei schwierigsten Wörter nicht mehr, so sinken die Korrelationen deutlich und fallen kleiner aus als jeweils die Zusammenhänge zwischen der Buchstaben- und Graphemzahl auf der einen und der Lösungshäufigkeit auf der anderen Seite. Insgesamt scheint die Itemstichprobe (18 Wörter) zu klein, um belastbare Aussagen bezüglich der Vorhersagekraft der von Naumann & Herné vorgeschlagenen Lupenstellenkategorien treffen zu können.

Ging es uns um die Analyse der von den Schülerinnen und Schülern gezeigten Kompetenz, so betrachten Naumann & Herné mit Blick auf die Itemkomplexität nicht mehr Daten auf Schülerebene. Vielmehr widmen sich die beiden Autoren der Frage, wie sich Itemschwierigkeiten vorhersagen lassen. Somit werden die Schwierigkeiten der Items zu den zu analysierenden Daten. Diagnostiziert werden nicht länger Kompetenzstände von Personengruppen oder Einzelpersonen, sondern Eigenschaften von *Testwörtern*. Unbestritten ist dies eine interessante Fragestellung, knüpft sie doch für den Bereich der Rechtschreibdiagnostik an die Tradition der Bestimmung von schwierigkeitsbestimmenden Merkmalen an, die in anderen Domänen bereits weit verbreitet und auch anerkannt ist (vgl. für das Hörverstehen Buck & Tatsuoka 1998). Allerdings bietet sich hier als methodischer Zugang weniger die Ermittlung korrelativer Zusammenhänge als vielmehr der Einsatz linearer Regressionsanalysen an, bei denen die Itemschwierigkeit als eine gewichtete Summe der einzelnen Lupenstellen des Wortes modelliert werden könnte (vgl. Hartig 2007).

1 Beide Befunde können von uns nicht exakt, sondern nur der Tendenz nach repliziert werden.

Was unter „Interaktion von Lupenstellen“ verstanden werden soll, erläutern Naumann & Herné zunächst anhand des folgenden Beispiels: „Schreibt ein Schüler z. B. <Wantafel>, so ist anzunehmen, dass eine morphologische Segmentierung in [vant] und [ta:fl] bzw. [ta:fəl] ihn leichter zur Ableitung des <d> im ersten Teil des Wortes geführt hätte.“ (9) Das wollen wir nicht bestreiten. Wir signieren diesen Fehler mit „MG“ („Morphemgrenze“) und nehmen wie Naumann & Herné an, dass der Schüler das Wort „Wand“ allein wohl nicht als <Wan> geschrieben hätte. Auf einen solchen Fehler wäre die Kategorie „MG“ ja auch gar nicht anwendbar. Insofern ist diese Schreibung für das, was Naumann & Herné als „Interaktion“ bezeichnen, u. E. als Beispiel nicht glücklich gewählt. Eine Interaktion im statistisch-methodischen Verständnis ist durch eine multiplikative Verknüpfung von Faktoren gekennzeichnet. Wenn bspw. zwei interessierende Bedingungen gemeinsam auftreten, dann beeinflussen sie sich wechselseitig und erzeugen einen Effekt, der sich nicht aus ihren jeweiligen Einzeleffekten – zum Beispiel durch eine bloße Addition derselben – ermitteln ließe. Was Naumann & Herné meinen, ist u. E. somit besser als *Kontextabhängigkeit* zu fassen.[2] Sie möchten betonen, dass die Schwierigkeit der Schreibung einer Lupenstelle *kontextbedingt* ist. So weise May darauf hin, dass „Verkehrsschild“ nur von einem Drittel, „Verkehr“ aber von etwa der Hälfte der von ihm getesteten Schülerinnen und Schüler richtig geschrieben wurde (10), einen analogen Befund präsentiere u. a. Scheele, die zeige, dass das <h> in „Fernsehprogramm“ von 17 Prozent, in „sieht“ von 11 Prozent und in „seht“ von lediglich vier Prozent der Probanden falsch geschrieben wurde (11). Auch Voss, Blatt & Kowalski (2007) werden in diesem Zusammenhang zitiert. Sie zeigen ebenfalls, dass Richtigschreibungen von Lupenstellen ein und derselben Kategorie kontextbedingt sein können: So ergibt sich für die Kategorie „Silbengelenk“ bei <kommen> eine Lösungshäufigkeit von 96 Prozent, bei <Schnurr-> aber nur von 34 Prozent, für „ß“ bei <süße> von 81 und bei <schließlich> von 57 Prozent, beim silbeninitialen „h“ bei <Reh> von 95 und bei <fröhlich> von 70 Prozent.

Voss, Blatt & Kowalski (2007) präsentieren aber noch andere Beispiele und Zahlen, auf die Naumann & Herné nicht eingehen: So kommen unter „v“ <vor> mit 92 Prozent und <Vorder> mit 68 Prozent, unter „f“ <Fuchs> mit 98 und <Fohlen> mit 50 Prozent, unter „Umlaut“ <nächsten> mit 84 und <glänzenden> mit 56 Prozent Lösungshäufigkeiten vor, schließlich unter „Dehnungs-h“ <ihre> mit 96 und <Fohlen> mit 58 Prozent. Wie lassen sich *diese* großen Differenzen erklären? Voss, Blatt & Kowalski (2007, 26) schreiben resümierend: „Zum einen ist die Aufgabenschwierigkeit ein und desselben Phänomens höher, wenn dieses in einer flektierten oder abgeleiteten Wortform vorkommt […]. (VI) Zum anderen wird ein und dasselbe Phänomen in bekannteren Wörtern häufiger richtig geschrieben als in unbekannten […].“

Dem Schluss, dass die Bekanntheit des (geschriebenen) Wortes ein relevanter Faktor ist, stimmen wohl auch Naumann & Herné zu. Ausdrücklich weisen sie ja auf das 2-Wege-Modell der Rechtschreibung hin. Im Rahmen eines bundesweit repräsentativen Tests kann man Testwörter im Hinblick auf dieses Merkmal allerdings nicht in

2 Unabhängig von der Begriffswahl wäre eine Prüfung statistischer Interaktionseffekte aus unserer Sicht interessant.

eine Rangfolge bringen. Eine Bedingung dafür wäre z. B., dass es in allen Bundesländern verbindliche Grundwortschätze gibt, deren Schnittmengen man ermitteln könnte. (VII) Das ist nicht der Fall.

Dass die Schwierigkeit einer Lupenstellenschreibung *auch* kontextbedingt sein kann, soll also nicht in Abrede gestellt werden. Relevant ist darüber hinaus der Aspekt, ob eine bestimmte Lupenstelle in einem häufigen und gut bekannten oder in einem für die Kinder unvertrauten Wort vorkommt.

Was die unterrichtliche Praxis angeht, so folgt u. E., was als didaktischer Konsens betrachtet werden kann: Soll nicht nur geschrieben werden, was das Herz begehrt, sondern sollen Schreibungen orthografisch „gesichert" werden, dann empfiehlt es sich, zunächst auf Wörter zu setzen, die strukturell eher einfach *und* häufig sind. Diese Häufigkeit wiederum beeinflusst die Schwierigkeit einer Wortschreibung und somit die Schwierigkeit der in ein Wort eingebetteten Lupenstellen.

Betrachtet man unsere Fehlerkategorien, so ist denkbar, dass sie zu einer systematischen Unterschätzung *wortspezifischen* Wissens verleiten. Denn sieht man von den Kategorien „Spezielle Grapheme", (partiell) „Vokallänge in der Minderheit der Fälle"[3] und von der Restkategorie „anderer Fehler" ab, dann zielen die übrigen Kategorien auf *Regelwissen*. Werden Wörter als Ganzheiten erinnert, dann ist die Bedingung der lokalen Unabhängigkeit der Fehlerlupenstellen innerhalb eines Wortes verletzt. Im Rahmen der Item-Response-Theorie versteht man darunter, dass die Wahrscheinlichkeit, ein bestimmtes Item zu lösen, nicht davon abhängig sein soll, dass man vorher ein anderes Item gelöst oder nicht gelöst hat. Denn diese Wahrscheinlichkeit soll ja nur von der Fähigkeit der Person und einem Itemparameter (der Schwierigkeit des Items) abhängen. Wenn Naumann & Herné die „Interaktion von Lupenstellen" bzw. – in unserer Terminologie – ihre *Kontextabhängigkeit* thematisieren, dann postulieren sie, dass die Unabhängigkeitsbedingung nicht gegeben ist. Allerdings betonen sie wie gezeigt nicht, dass *wortspezifisch* gelernt werde. Wir haben die „Ganzheitsthese" überprüft, indem wir jedes einzelne Wort, das die Kinder zu schreiben hatten, wie eine komplexe Aufgabe aus Teilen behandelt haben, deren Lösungen *nicht* unabhängig voneinander sind. Das Ergebnis zeigt: Was „unsere" Wörter angeht, so können die wortspezifischen Effekte vernachlässigt werden (Böhme & Bremerich-Vos 2009, 351). (VIII)

Sind diese Testwörter hinsichtlich ihrer Itemkomplexität und ihrer Kontextabhängigkeit bzw. Vertrautheit im Sinne von Naumnn & Herné *zu* schwierig? Dieser Vorwurf scheint bei den beiden Autoren mitzuschwingen, betonen sie doch, „dass komplexe Testwörter, offenbar aus Ökonomie- und/oder Normierungsgründen, gewissermaßen selbstverständlich geworden sind, was sich hier als begrenzt sinnvoll zeigt." (7) Es hat sich empirisch gezeigt, dass unser Test *nicht* zu schwierig war. Für Viert-

3 Bei der Vokallänge in der Minderheit der Fälle geht es um die Doppelschreibung von Vokalbuchstaben und vor allem um das Dehnungs-h. Für die Schreibung dieses Zeichens gibt es notwendige, aber keine hinreichenden Bedingungen.

klässler haben sich sogar leichte Deckeneffekte ergeben, d. h. es kann im unteren Leistungsbereich besser differenziert werden als im oberen.

4 Die dimensionale Struktur der Rechtschreibkompetenz

Wie im Rahmen großer Schulleistungsstudien üblich, haben wir uns bei unserer Testung auf die sogenannte Item-Response-Theorie gestützt. Das Testmodell ist probabilistischer Natur. In der einparametrischen Version wird angenommen, dass die Wahrscheinlichkeit, mit der eine Person ein Item löst, ausschließlich auf der Itemschwierigkeit und der Personenfähigkeit beruht. Die Personenfähigkeit bzw. Kompetenz ist nicht direkt beobachtbar, sondern muss als latente Größe erschlossen werden. Wird nach der Dimensionalität dieser Kompetenz gefragt, dann interessiert in erster Linie, ob es sich um eine einzige, homogene Größe handelt oder ob mehrere Teilkompetenzen unterschieden werden können. Die Erörterung dieser Frage ist aus unserer Sicht entscheidend und steht erst an ihrem Anfang (vgl. Böhme & Bremerich-Vos 2009, 346–353). Es zeigte sich, dass weder ein neundimensionales Modell, bei dem alle von uns angenommenen Fehlerkategorien als eigenständige Dimensionen verstanden werden, noch sechs- und dreidimensionale Modelle mit den Daten hinreichend übereinstimmten. Dass dafür zumindest teilweise auch Untersuchungsmängel verantwortlich sein könnten, die messmethodische Erschwernisse mit sich bringen, haben wir ausdrücklich eingeräumt. Weiterhin sollte berücksichtigt werden, dass für die Überprüfung der dimensionalen Struktur der Rechtschreibkompetenz zahlreiche methodische Zugänge denkbar sind, die im Rahmen unseres Beitrags noch keine Berücksichtigung gefunden haben. Hier kommen insbesondere nicht-parametrische Ansätze (DETECT, vgl. Stout 2002) in Betracht, die allerdings ebenfalls eher eine eindimensionale Modellierung des Konstrukts nahelegen (Böhme & Robitzsch 2009a). Somit halten wir unsere Schlussfolgerung, „dass näherungsweise mit einem eindimensionalen Modell zur Beschreibung der Rechtschreibkompetenz gearbeitet werden kann“ (Böhme & Bremerich-Vos 2009, 350), nach wie vor für plausibel.

Dass für die von uns untersuchte Altersgruppe auch hinsichtlich einer anderen, mit dem Deutschen eng verwandten Sprache eine ähnliche Ansicht vertreten wird, wollten wir mit dem Verweis auf die Studie von Notenboom & Reitsma (2003) belegen. Diese Autoren schlussfolgern für das Niederländische (nicht das Englische!):

> „In this study, the latent structure of a spelling achievement test for elementary school grades was investigated. Factor analyses revealed that for Grades 2 to 6, the test was unidimensional, whereas for Grade 1, two factors were found: a phonological factor and an orthographic factor. “ (1039) (IX)

Naumann & Herné weisen darauf hin, dass sich das Bild ändert, wenn man die Definitonen der Kategorien partiell modifiziert und ihre Zahl vergrößert, so dass fast doppelt so viele Lupenstellen resultieren. Aber auch wenn es sich so verhält, wird unsere These, dass ein eindimensionales Modell psychometrisch am plausibelsten sei, nicht obsolet. In der einschlägigen Literatur wurde nämlich verschiedentlich dis-

kutiert, dass höherkomplexe Modelle, bei denen eine große Zahl von Dimensionen angenommen wird, in der Regel eine schlechtere Passung aufweisen als Modelle geringerer Komplexität (vgl. Stout 2002 und Böhme & Robitzsch 2009b zur Frage essentieller Eindimensionalität, illustriert am Beispiel der Lesekompetenz).(X)

Zur Unterstützung ihrer Zweifel an der Eindimensionalitätsannahme zitieren Naumann & Herné aus der Arbeit von Voss, Blatt und Kowalski (2007, 25): „Aus analytischer Sicht kann das komplexere fünfdimensionale Kompetenzmodell die in den erfassten Schülerdaten enthaltenen Informationen besser darstellen als ein Generalfaktormodell […]." Irritiert halten sie aber auch folgenden Satz aus dieser Studie fest (ebd., 29): „Die Wortschreibungen im Kernbereich stellen ein und dieselbe Kompetenzleistung dar […]." Betrachtet man, was Voss, Blatt und Kowalski (2007) unter den Überschriften „Phonographisches und silbisches Prinzip" und „Morphologisches Prinzip" als Kernbereich fassen, und nimmt man einen Bereich hinzu, den sie als „Prinzipien der Wortbildung" bezeichnen, dann findet man hier, folgt man einigen Erläuterungen, außer Teilen der Groß- und Kleinschreibung und dem Dehnungs-h das, was auch uns interessiert. Die latenten, d. h. messfehlerbereinigten Korrelationen sind eindeutig: Für den phonographisch-silbischen und den morphologischen Kernbereich betragen sie .99, für Wortbildung und den phonographisch-silbischen Kernbereich .97 und für Wortbildung und den morphologischen Kernbereich .96 (Voss, Blatt & Kowalski 2007, 24).[4] Die Höhe dieser Korrelationen spricht eher dafür, dass eine Trennbarkeit der Teilfähigkeiten in verschiedene Dimensionen nicht gegeben ist, was Voss, Blatt & Kowalski selbst einräumen (25). Stützen lässt sich diese These durch eine Beobachtung, die wir im Kontext unserer Studie machen konnten: Ordnet man die Items den Dimensionen – also den Lupenstellenkategorien – nicht theoriegeleitet, sondern zufällig zu, so ergeben sich für unseren Itempool mittlere latente Korrelationen zwischen den Dimensionen in Höhe von $r = .95$ (vgl. Böhme & Robitzsch 2009a).

5 Fazit

Auch aus unserer Sicht ist interessant, was Naumann & Herné insbesondere zur Interaktion bzw. Kontextabhängigkeit von Lupenstellen und zur Komplexität von Items ausführen. Diese Überlegungen könnten die Entwicklung eines Modells der Rechtschreibkompetenz gewinnbringend ergänzen. Hierbei könnten – in kleinerem Rahmen – Zusatzstudien mit systematisch aufeinander bezogenen und variierten „einfachen" und „komplexen" Items weiterhelfen (vgl. Naumann & Herné, Fn. 9). Im Rahmen von „Großuntersuchungen" dürfte es aber nicht möglich sein, eine Korngröße vorzusehen, wie sie Naumann & Herné möglicherweise erstrebenswert finden, wenn sie Scheele zustimmend zitieren: „Die Einzelwortanalysen zu allen

4 Die Korrelationen mit einem Peripheriebereich fallen deutlich geringer aus. Neben dem Dehnungs-h geht es im Modell von Voss, Blatt & Kowalski (2007) hier vornehmlich um Verdopplungen des Vokalgraphems, nicht ableitbare einsilbige Wörter und Fremdwörter. Außer dem Dehnungs-h kamen Beispiele für diese Kategorien in unserer Testung nicht vor.

Fehlerkategorien haben […] ergeben, dass innerhalb jeder Kategorie weiter nach der Häufigkeit der Items bzw. Morpheme, den in einem Item zusätzlich vorhandenen Lupenstellen, den an den Lupenstellen betroffenen Graphemen, den Umgebungsbedingungen der Lupenstellen usw. unterschieden werden muss.“ (Naumann & Herné, 11)

Und die schulische Praxis? Wenn es zutreffen sollte, dass schwache Rechtschreiber in der Regel in *allen* Bereichen der orthografischen „Stufe“ Schwierigkeiten haben, dann schließt das aus unserer Sicht nicht aus, dass sie bereichs*spezifisch* gefördert werden sollten. So mag aus deklarativem Wissen um eine Regel in einem Bereich prozedurales Wissen werden, von dem man auch in anderen Bereichen profitieren kann.

Literatur

Böhme, Katrin & Bremerich-Vos, Albert (2009). Diagnostik der Rechtschreibkompetenz in der Grundschule – Konstruktprüfung mittels Fehler- und Dimensionsanalysen. In: D. Granzer, O. Köller, A. Bremerich-Vos u. a. (Hg.), Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule. Weinheim: Beltz, S.330–356.

Böhme, Katrin & Robitzsch, Alexander (2009a). Das Problem lokaler Abhängigkeiten von Items - Fehleranalysen in der Rechtschreibdiagnostik. Vortrag auf der 9. Fachtagung der Fachgruppe Methoden und Evaluation, Bielefeld, 09.–12. September 2009.

Böhme, Katrin & Robitzsch, Alexander (2009b). Methodische Aspekte der Erfassung der Lesekompetenz. In: D. Granzer, O. Köller, A. Bremerich-Vos u. a. (Hg.), Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule. Weinheim: Beltz, S.250–289.

Buck, Gary & Tatsuoka, Kikumi (1998). Application of the rule-space procedure to language testing: examining attributes of a free response listening test. In: Language Testing, 15 (2), 119–157.

Hartig, Johannes (2007). Skalierung und Definition von Kompetenzniveaus. In: B. Beck & E. Klieme (Hg.), Sprachliche Kompetenzen. Konzepte und Messung. Weinheim: Beltz, S. 83–99.

Herné, Karl-Ludwig & Naumann, Carl Ludwig (2002). Aachener Förderdiagnostische Rechtschreibfehler-Analyse (AFRA). Systematische Einführung in die Praxis der Fehleranalyse mit Auswertungshilfen zu insgesamt 33 standardisierten Testverfahren als Kopiervorlagen. Aachen: Alfa Zentaurus.

Naumann, Carl Ludwig & Herné, Karl-Ludwig (2010). Warum ist die Tomate leichter als das Fahrradschloss? Überlegungen zu Itemkomplexität und Kompetenzmodellierung in der Rechtschreibung. In diesem Heft.

Notenboom, Annelise & Reitsma, Pieter (2003). Investigating the dimensions of spelling ability. In: Educational and Psychological Measurement 63, 1039–1059.

Stout, William (2002). Psychometrics: From practice to theory and back: In: Psychometrika, 67, 485-518.

Voss, Andreas, Blatt, Inge & Kowalski, Kerstin (2007). Zur Erfassung orthographischer Kompetenz in IGLU 2006: Dargestellt an einem sprachsystematischen Test auf Grundlage von Daten aus der IGLU-Voruntersuchung. In: Didaktik Deutsch, Heft 23, 15–32.

Anschrift der Verfasser:

Katrin Böhme, Prof. Dr. Albert Bremerich-Vos, Fakultät für Geisteswissenschaften, Germanistik/Linguistik/Sprachdidaktik, Universität Duisburg-Essen, Universitätsstr. 12, 45117 Essen

katrin.boehme@uni-due.de; albert.bremerich-vos@uni-due.de

Carl Ludwig Naumann/Karl-Ludwig Herné

Statt einer Re-Replik - die vielen konsenten Punkte unerwähnt, aber sozusagen ein paar Kirschtomätchen nachgereicht

(I) Damit wäre unser Wunsch ‚erledigt' und verschiedene Auffassungen von ‚Rechtschreibkompetenz' pass(t)en dann nur bedingt zusammen, auch der Erwerbsprozess und sein Ergebnis. Wir stehen zu unserer Unzufriedenheit.

(II) Das belegt interne Stimmigkeit, sagt jedoch nichts zur Angemessenheit des Kategoriensystems.

(III) Aber Länge *und* Kürze ziehen graphische Besonderheiten nach sich, die, anders als die Umlautschreibung, z. T. nur durch Einprägen zu erlernen sind (vgl. die Gedächtnisarten in Abb. 1, Naumann & Herné, 8). Auch die von Böhme und Bremerich-Vos angeführte schulische „Standardlehre" geht u. W. nicht von einer Normalität der Kürze aus.

(IV) Das stellen wir gerade zur Debatte, vgl. Naumann & Herné, S. 10: „Zur Forschungslage".

(V) Zu ähnlichen Ergebnissen sind wir mit der Korrelation zwischen Phonemanzahl und Lösungshäufigkeiten gelangt. Es besteht kein statistisch signifikanter Zusammenhang. Stellt man zwei Wörter unterschiedlicher Itemkomplexität, aber annähernd gleicher Wortlänge, wie z. B. <Schlittschuhläufer> und <Vogelfutter>, einander gegenüber, dann lassen sich die unterschiedlichen Lösungshäufigkeiten nicht zur Wortlänge, wohl aber zur Itemkomplexität in Beziehung setzen (vgl. Naumann & Herné, Tab 2, 14).

(VI) Gerade das gerade sagen wir allerdings mehrfach.