Originalarbeit

WOR-TE: Ein Ja / Nein-Wortschatztest für Kinder verschiedener Altersgruppen

Entwicklung und Validierung basierend auf dem Rasch-Modell

Jutta Trautwein und Sascha Schroeder

Zusammenfassung: In dem vorliegenden Artikel wird der Wortschatztest WOR-TE für deutsche Grundschulkinder vorgestellt. Der Test basiert auf der Ja / Nein-Methode, bei der die Teilnehmerinnen und Teilnehmer aus einer Liste von Wörtern und Pseudowörtern diejenigen ankreuzen sollen, die sie kennen. Er wurde für verschiedene Altersgruppen (1./2. Klasse, 3./4. Klasse, 5./6. Klasse) konzipiert und Item-Response-Theory-basiert mit dem Rasch-Modell validiert. Anhand des Validitätskonzepts nach Messick (1995) wurden verschiedene Aspekte von Konstruktvalidität untersucht: Inhaltliche Aspekte, Relevanz, Repräsentativität, Technische Qualität, substantielle Aspekte, Generalisierbarkeit und externe Aspekte. Die Ergebnisse zeigen, dass der Test ein valides Instrument zur Erfassung des orthographischen Wortschatzes im Grundschulalter darstellt und durch kleine Veränderungen, insbesondere in Bezug auf die Item-Auswahl, profitieren würde. Möglichkeiten des Einsatzes sowie Einschränkungen des Verfahrens werden diskutiert, ebenso wie die Verwendung des Validitätskonzeptes mit IRT für Ja / Nein-Test im Allgemeinen.

Schlüsselwörter: Wortschatztest, Ja / Nein-Methode, Item Response Theory, Rasch Modell, Validierung

A *Yes / No* Vocabulary Test for Children of Different Age Groups: Development and Validation Based on the Rasch Model

Abstract: In this article we present a vocabulary test for German primary school children. The test is based on the *yes / no* method where participants identify words they know out of a list of words and pseudowords. The test was developed for different age groups (Grade 1/2, Grade 3/4, Grade 5/6) and validated via item response theory (IRT), namely, the Rasch model. Following the concept suggested by Messick (1995), we analyzed several aspects of construct validity: content aspects, relevance, representativity, technical quality, substantial aspects, generalizability, and external aspects. Results show that the test is a valid instrument for measuring the orthographic vocabulary of German primary school children but could also benefit from some minor changes concerning, for example, item selection. Possible applications and limitations of the instrument are discussed as well as the use of the validity concept and the validation via IRT for *yes / no* vocabulary tests in general.

Keywords: vocabulary test, *yes / no* method, item response theory, Rasch model, validation

Der Wortschatz stellt eine essenzielle Komponente der Sprachkompetenz dar und hängt eng mit der Lesefähigkeit und dem Schulerfolg zusammen (Biemiller, 2003, 2006). Er gilt daher in vielen frühen Screeningverfahren als Indikator für eine Sprachentwicklungsstörung (z. B. Elternfragebögen für die Früherkennung von Risikokindern – ELFRA; Grimm & Doil, 2006). Im späteren Spracherwerb wird der Wortschatz als Prädiktor für Lese- und Schreibfertigkeiten angesehen. So konnten Muter, Hulme, Snowling und Stevenson (2004) zeigen, dass die Wortschatzgröße zu Beginn der Grundschule das spätere Leseverständnis vorhersagte. Auch in querschnittlichen Untersuchungen wurde ein Zusammenhang zwischen Wortschatz und Leseverständnis nachgewiesen (Ricketts, Nation & Bishop, 2007; Ouellette & Beers, 2010). Zudem wird die Lesegeschwindigkeit als Teil der Lesefähigkeit vom Wortschatz beeinflusst (z. B. Anderson, Wilson & Fielding, 1988). Perfetti und Stafura (2013) nehmen an, dass ein besserer Wortschatz den lexikalischen Zugriff erleichtert, was wiederum das Leseverständnis begünstigt. Da die Lesefähigkeit zentral für den Schul- und daran anschließenden beruflichen Erfolg sowie die Teilhabe an der Gesellschaft ist, ist es entscheidend, Defizite und ihre Gründe früh aufzudecken, um effektive Trainingsmethoden einzuleiten (Biemiller, 2005).

https://doi.org/10.1026/0012-1924/a000212

Das Konstrukt *Wortschatz* ist nicht leicht zu definieren. Perfetti und Hart (2002) gehen in ihrer Hypothese der lexikalischen Qualität davon aus, dass der Wortschatz eine phonologische, eine orthographische und eine semantische Komponente umfasst. Die phonologische Komponente beinhaltet Wissen über die Aussprache, die orthographische über die Schreibung und die semantische über die Bedeutung eines Wortes. Die verschiedenen Wissensebenen können für ein Wort unterschiedlich stark ausgeprägt sein. Beim Lesen muss demnach zunächst die orthographische Form des Wortes abgerufen werden, für das Leseverständnis zudem das semantische Wissen und zum lauten Lesen die phonologische Komponente. So wird es auch häufig in Modellen zur visuellen Worterkennung angenommen (z. B. das Dual Route Model; Coltheart, Rastle, Perry, Lagdon & Ziegler, 2001). Demnach ist insbesondere der orthographische Wortschatz für die Lesefähigkeit entscheidend.

Eine Möglichkeit zur Messung des orthographischen Wortschatzes ist die Ja / Nein-Methode von Anderson und Freebody (1983). Teilnehmende identifizieren alle ihnen bekannten Wörter innerhalb einer Wortliste. Um Raten zu vermeiden, enthält die Liste auch Pseudowörter. Anderson und Freebody (1983) fanden bei Fünftklässlerinnen und Fünftklässlern hohe Korrelationen mit mündlichen Definitionsaufgaben und Multiple-Choice-Wortschatztests und den Ergebnissen aus dem Ja / Nein-Test (r = .84 für Multiple Choice, r > .85 für Definitionsaufgaben). Die Autoren haben demnach ein valides Instrument zur Erfassung des orthographischen Wortschatzes entwickelt und zudem Zusammenhänge zu semantischem Wissen über Wörter gefunden. Ähnliche Ergebnisse erzielten auch andere Studien (z. B. Mochida & Harrington, 2006: r = .85 für Multiple Choice; Pellicer-Sánchez & Schmitt, 2012: r = .89 für mündliche Definitionen). Obwohl Anderson und Freebody (1983) den Wert von Ja / Nein-Tests auch für Kinder demonstriert haben, wird er bislang vornehmlich für die Messung des Wortschatzes von Erwachsenen in einer Zweitsprache verwendet (z. B. Eyckmans, 2004; Huibregtse, Admiraal & Merea, 2002; Merea & Buxton, 1987; Mochida & Harrington, 2006; Lemhöfer & Broersma, 2012). Das Testformat unterscheidet sich von anderen bereits bestehenden Verfahren für Kinder im Deutschen, welche vielmals primär auf die semantische Ebene des Wortschatzes abzielen (z. B. *Peabody Picture Vocabulary Test* – PPVT-4; Lenhard, Lenhard, Segerer & Suggate, 2015; Patholinguistische Diagnostik bei Sprachentwicklungsstörungen – PDSS, Kauschke & Siegmüller, 2009). Hierbei wird häufig mit dem Benennen oder Zeigen von Bildern nach mündlicher Vorgabe gearbeitet. Zudem sind existierende Verfahren meist nur in Einzelerhebungen durchführbar und zielen oftmals auf die Diagnose semantischer Defizite ab (z. B. *Wortschatz- und Wortfindungstest* – WWT; Glück, 2011). Die meisten dieser bereits existierenden Instrumente sind außerdem für Vorschulkinder konzipiert worden. Ein weiteres Testformat für Schulkinder und Erwachsene ist die Auswahl von Synonymen nach schriftlicher Vorgabe (z. B. *Grundintelligenztest* – CFT-20; Weiß, 2006). Allerdings ist die Aufgabe sehr stark von den Distraktor-Items abhängig. Das Wissen über die Bedeutung der Distraktor-Items kann demnach die Lösung der Aufgabe beeinflussen (Anderson & Freebody, 1983). Mit steigendem Alter wird dies aber schwieriger, da der Wortschatz substanziell wächst, insbesondere im Schulalter (Segbers & Schroeder, 2017; Anglin, Miller & Wakefield, 1993). Geeignete Verfahren für Schulkinder, die für verschiedene Altersgruppen geeignet sind, gibt es im Deutschen aktuell nicht.

Da die Ergebnisse zum Ja / Nein-Testformat von Anderson und Freebody (1983) ermutigend sind und die Notwendigkeit besteht, den orthographischen Wortschatz im Grundschulalter zu messen, ist es vielversprechend, den Tests für das Grundschulalter zu adaptieren. Tatsächlich birgt das Verfahren Vorteile für die praktische Anwendung: Wegen des geringen kognitiven Aufwands kann eine große Anzahl an Items in kurzer Zeit dargeboten werden. Zudem kann der Test in Gruppentestungen durchgeführt werden.

Bisher wurden die Verfahren lediglich über die Korrelationen zu Definitionsaufgaben oder Multiple-Choice-Verfahren validiert (z. B. Anderson & Freebody, 1983; Mochida & Harrington, 2006; Pellicer-Sánchez & Schmitt, 2012). Messick (1995) zufolge bezieht sich diese Art von Validierung auf die konvergente Validität, welche durch die Korrelation des Testscores mit externen Variablen, die dasselbe oder assoziierte Konstrukte messen, definiert ist. Er nennt allerdings noch weitere Aspekte, die zur Validierung eines Tests herangezogen werden sollten. Messick (1995) beschreibt inhaltliche Aspekte, die die Relevanz, die Repräsentativität und die technische Qualität des Testinhalts umfassen. Sie zielen damit darauf ab, zu überprüfen, inwiefern die Inhalte eines Tests zur Messung der entsprechenden Fähigkeit angemessen sind. Er beschreibt auch substanzielle Aspekte, die sich auf die Einbettung der Testergebnisse in ein nomologisches Netzwerk beziehen. Damit ist die Passung der Testergebnisse zu vorherigen Annahmen in Bezug auf die gemessene Fähigkeit gemeint. Des Weiteren nennt er strukturelle Aspekte, die sich auf Annahmen zur Struktur des zu messenden Konstrukts beziehen, Generalisierbarkeit, die die Adaption des Testformats für andere Items oder andere Teilnehmende meint, und externe Aspekte, die die Korrelation mit konvergenten und divergenten Variablen beinhaltet. Ein Ansatz zur Anwendung dieses Konzepts der Validierung für Wortschatztests wurde von Beglar (2010) sowie McLean, Kramer und Beglar (2015) vorgestellt. Sie untersuchten die verschiedenen Aspekte von Validität

Tabelle 1. Verteilung der Frequenz und Itemschwierigkeit auf die drei Testversionen

	Log Lemma Frequenz		Itemschwierigkeit
Testversion	*M* (*SD*)	Bereich	*M* (*SD*)
Klasse 1/2	1.5 (0.4)	2.7 – 1.0	-0.62 (0.88)
Klasse 3/4	0.6 (0.1)	0.9 – 0.4	1.05 (1.38)
Klasse 5/6	0.0 (0.1)	0.4 – 0.2	2.56 (1.46)

unter der Verwendung der Item Response Theory (IRT) anhand des Rasch-Modells. Shillaw (1996) zeigte zudem bereits, dass das Rasch-Modell für die Auswertung von Ja / Nein-Wortschatztests geeignet ist.

In der vorliegenden Studie wird ein Ja / Nein-Wortschatztest *WOR-TE* für deutsche Grundschulkinder verschiedener Altersgruppen vorgestellt. Beruhend auf dem Konzept der Validität von Messick (1995) sowie dem IRT-basierten Ansatz von Beglar (2010) und McLean et al. (2015) soll dabei gezeigt werden, dass es sich bei dem Test um ein valides Instrument zur Erfassung des orthographischen Wortschatzes von Grundschulkindern handelt. Dazu werden die Testergebnisse mithilfe des Rasch-Modells skaliert und auf die verschiedenen Aspekte der Validität nach Messick (1995) untersucht.

Testentwicklung

Um eine breite Altersspanne von Grundschulkindern abzudecken, wurden drei Testversionen des WOR-TE (Wortschatz-Test) für verschiedene Altersgruppen (1./2. Klasse, 3./4. Klasse, 5./6. Klasse) entwickelt. Da die Wortfrequenz die Itemschwierigkeit in Wortschatztests hauptsächlich bestimmt (z. B. Beglar, 2010), wurde die mittlere Lemmafrequenz[1] der Items (childLex Kinderkorpus; Schroeder, Würzner, Heister, Geyken & Kliegl, 2015) in den verschiedenen Testversionen systematisch manipuliert (Tabelle 1). Die Auswahl passender Frequenzen für jede Altersgruppe basierte auf Ergebnissen von vorherigen Studien (u. a. Developmental Lexicon Study; Schröter & Schroeder, 2017). Die Materialien beinhalteten Nomen, Verben und Adjektive.

Jede Testversion umfasste 100 Wörter. Um einen Vergleich der drei Testversionen zu ermöglichen, waren 20 Wörter in allen Testversionen identisch. Diese 20 Anker-Items wurden aus dem Frequenzbereich von allen drei Testversionen ausgewählt. Zusätzlich teilten sich aufeinanderfolgende Testversionen jeweils zehn Items. Das bedeutet, die Testversion für die 1. und 2. Klasse umfasste 70 unique Items, 20 Anker-Items, die in allen Testversionen enthalten waren, und 10 Anker-Items, die ebenfalls in der Version für die 3. und 4. Klasse enthalten waren. Die Testversion für die 3. und 4. Klasse enthielt 60 unique Items, die 20 allgemeinen Anker-Items, 10 geteilte Anker-Items mit der 1. und 2. Klasse und 10 geteilte Anker-Items mit der Version für die 5. und 6. Klasse. Die Testversion für die 5. und 6. Klasse enthielt 70 unique Items, die 20 allgemeinen Anker-Items sowie die 10 geteilten Items aus der Version für die 3. und 4. Klasse. Der Test umfasst damit insgesamt 240 Items.

Um das Raten zu minimieren, wurden zu jeder Testversion 24 Pseudowörter hinzugefügt. Diese wurden durch das Austauschen von mindestens einem Buchstaben in einem realen Wort bzw. die Aneinanderreihung von Morphemen konstruiert und waren in jeder Testversion gleich. Für jede Altersgruppe wurden zwei Pseudoparallel-Versionen A und B mit randomisierter Item-Reihenfolge erstellt.

Die wortwörtliche Instruktion für die teilnehmenden Kinder lautete: „Im Folgenden seht ihr eine Liste von Wörtern. Ihr sollt die Wörter markieren, die ihr kennt. Dabei dürft ihr nicht raten, denn die Liste enthält auch Wörter, die es gar nicht gibt. Wenn ihr ratet, merken wir das sofort. Kreuzt nur die Wörter an, die ihr wirklich kennt." Drei Beispielitems (2 Wörter und 1 Pseudowort) wurden zur Veranschaulichung der Aufgabe besprochen. Abhängig von der Altersgruppe dauerte die Testdurchführung 5 bis 15 Minuten.

Methode

Stichprobe

Insgesamt nahmen $N = 422$ Kinder (Klassen 1–6) von fünf Berliner Grundschulen an der Studie teil. Vierundzwanzig Kinder (6 %) füllten den Wortschatztest unvollständig aus und wurden daher aus den weiteren Analysen ausgeschlossen, sodass die Daten von $N = 398$ Kindern (198

[1] Als Lemma wird die zitierfähige Grundform eines Wortes bezeichnet.

Tabelle 2. Stichprobenbeschreibung und mittlere Hit- und False-Alarm-Raten pro Altersgruppe

			Geschlecht			Muttersprache				
Klasse	*N*	*M* Alter (*SD*)	männl.	weibl.	NA	D	ND	NA	*M* Hit Rate (*SD*)	*M* False Alarm Rate (*SD*)
1	37	6.6 (0.5)	13	23	1	18	16	3	.56 (.17)	.16 (.10)
2	49	7.3 (0.7)	24	25	0	33	14	2	.53 (.20)	.11 (.11)
3	75	8.0 (0.6)	38	35	2	49	23	3	.38 (.18)	.05 (.07)
4	107	9.0 (0.6)	65	42	0	67	37	3	.52 (.17)	.05 (.07)
5	62	10.0 (0.6)	22	40	0	35	26	1	.30 (.14)	.05 (.05)
6	68	11.2 (0.6)	35	33	0	31	34	3	.44 (.16)	.05 (.07)
total	398	8.9 (1.6)	197	198	3	233	150	15	.45 (.19)	.06 (.08)

Anmerkungen: NA = Keine Angabe, D = Deutsch als einzige Muttersprache, ND = weitere Muttersprachen neben Deutsch.

weiblich, 197 männlich, 3 ohne Geschlechterangabe) verwendet werden konnten. Ein Großteil der Kinder (233, 59 %) gab Deutsch als ihre einzige Muttersprache an, während 150 Kinder (38 %) angaben, mindestens eine weitere Muttersprache gelernt zu haben. Eigenschaften der Stichprobe sind in Tabelle 2 enthalten.

Instrumente

Der Wortschatz wurde mit dem Subtest Sprachverständnis (*Kognitiver Fähigkeitstest* – KFT1–3, Heller & Geisler, 1983) bzw. Wortschatz (*Kognitiver Fähigkeitstest* – KFT 4–12+ R, Heller & Perleth, 2000) untersucht. In der Version für die 1. bis 3. Klassenstufe wählen die Kinder nach auditiver Vorgabe ein passendes Bild aus fünf Bildern aus. Für die 4. bis 12. Klasse handelt es sich um ein Multiple-Choice-Verfahren, wobei zu einem fettgedruckten Wort das passende Synonym gesucht werden muss. Die Rohwerte wurden in jahrgangsspezifische T-Werte überführt. Die Reliabilität wurde mit Cronbachs α von .57 (1.–3. Klasse) bzw. .71 (4.–6. Klasse) bestimmt. Für die früheren Klassen ist sie damit zu gering, in den höheren Altersstufen akzeptabel.

Die Lesegeschwindigkeit wurde mit dem *Salzburger Lese-Screening für die Klassenstufe 1–4* – SLS 1–4 (Mayringer & Wimmer, 2003) bzw. 5–8 (Auer, Gruber, Mayringer & Wimmer, 2005) erfasst. Dabei sollen die Kinder innerhalb von drei Minuten für möglichst viele Sätze angeben, ob sie wahr oder falsch sind. Der Testscore ergibt sich aus den korrekt markierten Sätzen. Es werden alterskorrigierte Normwerte verwendet. Cronbachs α zur Überprüfung der Reliabilität lag bei .96 und ist somit sehr gut.

Orthographische Fähigkeiten wurden mit der *Hamburger Schreibprobe 1–9* (May, 2002) ermittelt. Dabei werden Wörter und Sätze diktiert und anschließend die richtigen Grapheme gezählt. Die Ergebnisse werden als alterskorrigierte T-Werte berichtet. Zur Berechnung der Reliabilität wurde die Anzahl richtiger Grapheme pro Wort verwendet. Da verschiedene Wörter pro Altersgruppe eingesetzt werden, wurde Crohnbachs α separat berechnet. Der Mittelwert war mit $M = .81$ sehr zufriedenstellend. Da die orthographischen Fähigkeiten in der 1. Klasse noch sehr stark schwanken, fand hier keine Erfassung statt.

Die nonverbale Intelligenz der Teilnehmenden wurde mit dem Matrizen-Subtest des CFT 1 (Cattell, Weiß & Osterland, 1997) bzw. CFT 20-R (Weiß, 2006) erhoben. Die Aufgaben bestehen jeweils aus einem Muster, welches mithilfe einer Auswahl von fünf Möglichkeiten vervollständigt werden muss. Testteilnehmerinnen und Testteilnehmer haben dafür sechs (CFT 1 für die Erstklässler) bzw. drei Minuten (CFT 20-R, ab Klasse 2) Zeit. Da lediglich ein Subtest durchgeführt wurde, können nur die Rohwerte (Anzahl richtiger Antworten) für die Analyse verwendet werden. Für die Überprüfung der Reliabilität wurde ein zufriedenstellender Wert von Cronbachs α mit .81 (1. Klasse) bzw. .68 (Klasse 2–6) berechnet.

Prozedur

Das schriftliche Einverständnis der Eltern war notwendig für die Studienteilnahme. Alle Aufgaben wurden während der Schulzeit innerhalb von zwei Schulstunden (à 45 Minuten) im Klassenverband durchgeführt. Zusätzlich wurden demographische Daten (Alter, Geschlecht und Muttersprache) mit einem Fragebogen ermittelt. Mithilfe von Identifikationsnummern wurden die Daten anonymisiert. Für die Teilnahme erhielten die Kinder ein kleines Dankeschön.

Analysen

Zur Analyse wurde eine Item-Response-Analyse unter Einsatz des Rasch-Modells (Embretson & Reise, 2000)

durchgeführt. Um Unterschiede zwischen den Altersgruppen zu berücksichtigen, wurden ein Multiple-Group-Modell gewählt, bei dem die verschiedenen Altersgruppen als separate Gruppen behandelt wurden (Bock & Zimowski, 1997). Die 20 Ankeritems ermöglichten dabei eine Schätzung der Itemparameter von allen Testversionen auf einer gemeinsamen Skala (Embretson & Reise, 2000). Die Modelle wurden mit dem TAM Paket für R (Kiefer, Robitzsch & Wu, 2016) geschätzt, welches Marginal Maximum Likelihood (MML) für die Parameterschätzung verwendet (Mislevy & Stocking, 1989). Für die Modellschätzung wurde *vertical linking* und *concurrent calibration* genutzt (für einen Überblick über Skalierungsmethoden siehe Kolen & Brennan, 2004). Die Modelle wurden identifiziert, indem der erste Itemparameter auf 0 fixiert wurde. Personenparameter, die das latente Personenmerkmal des Wortschatzes repräsentieren, wurden ebenfalls mit MML geschätzt. Aufgrund fehlender korrekter Antworten musste ein Item (*äsen*, Version 5./6. Klasse) zuvor ausgeschlossen werden. Auf die Prüfung der Modellpassung wird im Ergebnisteil eingegangen.

Ergebnisse

Die Raten der Hits und False Alarms sind in Tabelle 2 dargestellt. Im Folgenden werden in Bezug auf den Ja / Nein-Wortschatztest fünf verschiedene Aspekte von Konstruktvalidität nach Messick (1995) in Betracht gezogen. Im Anschluss werden die Passung des Rasch-Modells und die Validität des Tests bewertet und die Nützlichkeit der IRT-basierten Validierung diskutiert.

Inhaltliche Aspekte

Zunächst wurde überprüft, inwieweit der Inhalt des Ja / Nein-Wortschatztests angemessen ist, um den Wortschatz der Testteilnehmenden zu messen.

Inhaltliche Relevanz. Messick (1995) definiert die inhaltliche Relevanz als eine Auswahl von Aufgaben, die relevant für die Messung des Konstruktes sind. Für den vorliegenden Ja / Nein-Wortschatztest ist dies dadurch gegeben, dass die Wörter aus einem spezifischen Korpus für Kindersprache ausgewählt wurden. Diese Wörter werden daher mit hoher Wahrscheinlichkeit von den Kindern im Alltag rezipiert. Zur Anpassung an die jeweiligen Altersgruppen wurde zudem die Frequenz der Wörter für die verschiedenen Testversionen systematisch manipuliert (Tabelle 1). Zusätzlich dienten vorherige Studien (u. a. DeveL; Schröter & Schroeder, 2017) dazu, Wörter auszuwählen, die eine ausreichende Variabilität in den Erkennensraten bei der Zielgruppe hatten. Die inhaltliche Relevanz ist somit durch die Testkonstruktion gegeben.

Repräsentativität. Messick (1995) betont, dass ein Test alle wichtigen Teile des Konstrukts enthalten muss, um repräsentativ zu sein. Dies beinhaltet eine ausreichende Anzahl von Items, eine adäquate Streuung der Itemschwierigkeit und das Fehlen von Lücken in der Item-Hierarchie (Beglar, 2010). Abbildung 1 zeigt eine Item-Personen-Zuordnung für die Itemschwierigkeit und den Personenparameter aus den Testergebnissen. Links ist die Verteilung der Itemschwierigkeit dargestellt. Die rechte Seite repräsentiert die Verteilung der Personenparameter. Bezüglich der Anzahl von Items empfiehlt Beglar (2010) zehn Items pro Schwierigkeitsstufe. In Abbildung 1 ist zu sehen, dass dieses Kriterium für die meisten Schwierigkeitsstufen erfüllt wurde, lediglich an den Rändern der Verteilung ist die Anzahl etwas geringer. Der Test würde also profitieren, wenn man besonders einfache und besonders schwere Items hinzufügt. Die Streuung der Itemschwierigkeit erscheint ausreichend. Sie rangiert zwischen -2.95 und 5.37, wobei 97 % der Personenparameter zwischen -2.5 und 5 lag. Es können keine Lücken in der Itemhierarchie beobachtet werden. Allerdings können vier Items aus der Version für die 5./6. Klasse als zu schwer angesehen werden (*brüsk:* 5.37; *süffisant:* 5.05; *Häme:* 4.59; *schartig:* 4.59). In einer neuen Testversion sollten diese Items ausgelassen bzw. ersetzt werden. Die Verteilung der Personenparameter zeigt keinen Boden- oder Deckeneffekt und der mittlere Standardfehler SE = .03 (SD = .005) lässt auf eine präzise Messung der Personenfähigkeit schließen.

Um zusätzlich zu überprüfen, ob die Items repräsentativ für den gesamten Korpus sind, wurden die drei Maße Lemmafrequenz, Anzahl der orthographischen Nachbarn und Wortlänge der Items mit denjenigen der Wörter aus dem gesamten Korpus verglichen. Dafür wurden diejenigen Wörter aus dem gesamten Korpus herausgenommen, die nur einmal vorkamen unter der Annahme, dass sie nicht ausreichend repräsentativ für den Wortschatz eines Sprechers der Sprache sind. Der Vergleich der Maße ergab keinen Unterschied in der mittleren Frequenz der Items und der Wörter aus dem Gesamtkorpus, $t(239) = -0.10$, $p = .92$. Die Wortlänge unterschied sich dahingehend, dass im Gesamtkorpus insgesamt mehr längere Wörter enthalten waren, $t(239) = -19.80$, $p < .001$. Dies lässt sich dadurch erklären, dass im Gesamtkorpus auch Komposita enthalten sind, die für die Auswahl der Items nicht beachtet wurden. Bedingt durch die Wortlänge, die mit der Anzahl der orthographischen Nachbarn zusammenhängt, ergab sich auch bezüglich dieses Merkmals ein signifikanter Unterschied, $t(239) = 3.71$, $p < .001$. Schränkt man allerdings die Länge der Wörter im Gesamtkorpus entsprechend der Länge der Items ein, verschwindet die-

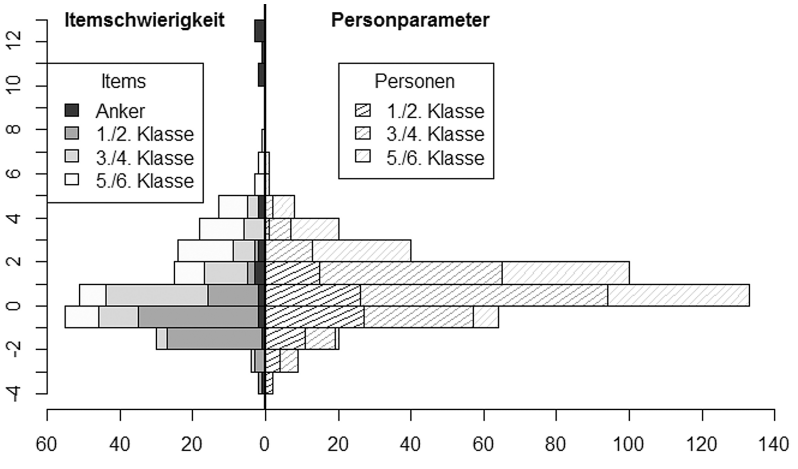

Abbildung 1. Verteilung von Personen- und Itemparametern für den WOR-TE.

se Differenz, t (239) = 1.10, p = 0.27. Insgesamt kann man davon ausgehen, dass die ausgewählten Items repräsentativ für den Korpus und damit für die verwendete Sprache in deutscher Kinderliteratur sind.

Technische Qualität. Technische Qualität meint die Passung der Items zum verwendeten Modell (Beglar, 2010). Um diese zu messen, wurde der Itemfit zum Rasch-Modell mit dem Maß der *Rasch infit mean-square statistic (MNSQ)*[2] bestimmt. Angelehnt an McNamara (1996) wurde ein Kriterium von ±2 Standardabweichungen vom Mittelwert der Infit Statistik (= 1) gewählt, um eine fehlende Passung zu identifizieren. Da die Standardabweichung 0.12 betrug, wird ein Infit-Wert zwischen 0.76 und 1.24 als akzeptabel für die Item-Passung behandelt. Kein Infit-Wert kleiner als 0.76 kann beobachtet werden, jedoch wird für zehn Items (4 %) ein zu hoher Wert gemessen (1./2. Klasse: *Planet*, Infit MNSQ = 1.38; *Backe*, Infit MNSQ = 1.31; *Statue*, Infit MNSQ = 1.31; *starren*, Infit MNSQ = 1.27; *passieren*, Infit MNSQ = 1.25; 5./6. Klasse: *sengen*, Infit MNSQ = 1.46; *Spind*, Infit MNSQ = 1.30; Ankeritems: *Tresse*, Infit MNSQ = 1.53; *wähnen*, Infit MNSQ = 1.34; *konstatieren*, Infit MNSQ = 1.30). Nach Ausschluss der unpassenden Items wurde ein neues Modell geschätzt und analog erneut der Itemfit überprüft. Dieser Prozess wurde so häufig wiederholt, bis keine unpassenden Items mehr vorhanden waren. Dabei wurden weitere 16 Items ausgeschlossen (1./2. Klasse: *Schüssel, Museum, knirschen, Strahl, ignorieren, reagieren, Stapel, Gegend, grell;* 3./4. Klasse: *artig*; 5./6. Klasse: *Galosche, Fanfare, schwelen, graziös, imitieren*; Ankeritems: *Reuse*).

Zusammenfassend lässt sich bezüglich der inhaltlichen Aspekte festhalten, dass die Items relevant und repräsentativ für das zu messende Konstrukt des orthographischen Wortschatzes sind, bis auf vier zu schwere Items. Durch die Analyse der technischen Qualität wurden zehn nicht passende Items identifiziert. Kleine Veränderungen an der Zusammensetzung der Items könnten zur Verbesserung des Tests beitragen. In einem nächsten Schritt wurden die unpassenden Items entfernt und ein neues Modell mit den 209 verbleibenden Items wurde geschätzt. Die neu geschätzten Parameter korrelierten hoch mit denen aus dem vorherigen Modell (Items: r = .99; Personen: r = .99). Die generellen Ergebnisse wurden durch die neue Modellschätzung also nicht verändert. Für die folgenden Analyseschritte wurde das Modell mit der reduzierten Itemanzahl verwendet. Um die Passung des Modells zusätzlich zu überprüfen, wurde zudem mithilfe der Q3-Statistik (Yen, 1984) evaluiert, ob die Antworten auf die Items unabhängig voneinander sind. Dazu wird die Residualkorrelation für alle Itempaare berechnet. Sie sollte bei lokaler Unabhängigkeit ungefähr 0 betragen. Im vorliegenden Modell ist dies der Fall, der Mittelwert der Q3-Statistik lag bei M = -0.01, SD = 0.09. Lediglich 3 % der gesamten Residualkorrelationen weichen mehr als zwei Standardabweichungen vom Mittelwert ab und können damit als Ausreißer angesehen werden. Die lokale Unabhängigkeit der Antworten ist damit gegeben, was zusätzliche Evidenz für die Passung des Rasch-Modells auf die Daten darstellt.

Substanzielle Aspekte

Der substanzielle Aspekt von Validität betrifft die Passung der Testergebnisse zu vorherigen Theorien in Bezug auf Prozesse, die die Testleistung beeinflussen (Messick,

[2] Mean Squares geben die χ^2 Statistik geteilt durch deren Freiheitsgerade an und zeigen inwiefern die tatsächlichen Antworten mit denen des Modells übereinstimmen.

1995). Es ist bereits bekannt, dass Wortfrequenz die Leistung in Aufgaben bei Wortschatztest beeinflusst. Um dies für die vorliegenden Daten zu überprüfen, wurde die Itemschwierigkeit mit der logarithmierten Lemmafrequenz der Items (childLex, Schroeder et al., 2015) korreliert. Mit r = -.74 kann die Korrelation als sehr hoch bezeichnet werden, was zeigt, dass hochfrequente Wörter einfacher zu erkennen sind.

In Bezug auf die Testergebnisse der Kinder wurden Altersgruppe, Geschlecht und Muttersprache als relevante Faktoren, die die Testleistung beeinflussen können, betrachtet. Um den Effekt dieser drei Variablen zu messen, wurde jeweils eine einfaktorielle ANOVA mit dem Personenparameter als abhängige Variable und Altersgruppe, Geschlecht, Muttersprache als unabhängige Variablen gerechnet. Dazu wurden die Personen, die keine Angaben zu Geschlecht oder Muttersprache gemacht haben, ausgeschlossen (N = 11). Die Modelle zu Muttersprache und Geschlecht enthielten zusätzlich das Alter (zentriert an der jeweiligen Altersgruppe) als Kontrollvariable. Die Ergebnisse zeigen einen signifikanten Effekt der Altersgruppe, $F(2,378) = 57.77$, $p < .001$, $\eta^2 = .27$. Post-hoc Analysen zeigen einen signifikanten Unterschied zwischen allen Altersgruppen, alle $p < .001$. Dies entspricht vorherigen Studien, da der Wortschatz mit dem Alter ansteigt (Segbers & Schroeder, 2017). Es gibt keinen Effekt des Geschlechts auf den orthographischen Wortschatz, $F(1, 378) = 2.3$, $p = .103$. Dies passt zu Ergebnissen aus vorherigen Studien, die keinen Unterschied im Wortschatz zwischen Jungen und Mädchen im Grundschulalter beobachtet haben (z. B. Anglin, Miller & Wakefield, 1993). Der Effekt der Muttersprache ist signifikant, $F(1, 378) = 6.15$, $p = .013$, $\eta^2 = .02$, und zeigt, dass monolinguale Kinder ein signifikant höheres Testergebnis hatten als bi- und multilinguale Kinder. Dies steht im Einklang mit früheren Ergebnissen zu besserer Wortschatzfähigkeit von monolingualen im Gegensatz zu bilingualen Personen (z. B. Bialystok, Luk, Peets & Yang, 2010).

Als ein weiterer Indikator für substanzielle Validität wurde das Rateverhalten betrachtet. Das Ja / Nein-Testformat beinhaltet Pseudowörter, um Rateverhalten zu minimieren. Für jedes Kind wurde der Anteil falsch ausgewählter Pseudowörter berechnet, um den Zusammenhang zwischen dem Rateverhalten und dem Testwert zu ermitteln. Der durchschnittliche Anteil angekreuzter Pseudwörter lag bei M = .07 (SD = .09). Für die einzelnen Klassenstufen ist sie in Tabelle 2 dargestellt. Sie ist damit in allen Klassen sehr niedrig, lediglich in der 1. und 2. Klasse war sie ein wenig erhöht. Allgemein raten Kinder bei der Durchführung nicht, was die allgemeine Konstruktvalidität des Instruments unterstützt. Zudem korreliert die False-Alarm-Rate nur mit r = -.15, t = -3.1, p = .002, mit dem Personenparameter aus dem Modell. Die Testleistung ist also weitgehend unabhängig von dem Antwortverhalten bei den Pseudowörtern. Der Einbezug eines Rateparameters in das Modell erscheint damit nicht indiziert. Im Diskussionsteil wird dieser Punkt noch einmal aufgegriffen.

Zusammenfassend ist festzuhalten, dass die Testergebnisse zu substanziellen Theorien auf der Item-Ebene bezüglich Korrelationen mit Wortfrequenz und auf der Personen-Ebene bezüglich des Einflusses von Alter, Geschlecht und Muttersprache passen. Zudem wird das Rateverhalten durch den Einbezug von Pseudowörtern erfolgreich minimiert.

Strukturelle Aspekte

Laut Messick (1995) ist es für die Testvalidität entscheidend, dass der Inhalt des Tests ein zugrundeliegendes Konzept misst. Im Fall des Ja / Nein-Tests handelt es sich hierbei um den orthographischen Wortschatz, der ein Konstrukt bzw. eine Dimension darstellt. Um zu überprüfen, ob der Test tatsächlich nur diese eine Dimension misst, wurde das Modell auf Eindimensionalität getestet. Dazu wurden zwei Modelle mit verschiedenen Dimensionen geschätzt und jeweils mit dem eindimensionalen anhand des Log Likelihoods verglichen. Im ersten Vergleichsmodell wurden die Dimensionen durch die drei Testversionen für verschiedene Altersgruppen definiert, sodass spezifische Items einer Testversion auf eine Dimension abgebildet wurden. Die Analyse ergab keinen signifikanten Unterschied zwischen den beiden Modellen, p = 1. In einem zweiten Vergleichsmodell wurden die Dimensionen nach Wortarten (Nomen, Verben, Adjektive) definiert, sodass jede Wortart eine Dimension darstellte. Auch hier zeigte sich kein signifikanter Unterschied zwischen den Modellen, p = 1.

Daraus lässt sich schlussfolgern, dass die Hinzunahme weiterer Dimensionen das Modell nicht verbessert, was die Annahme von Eindimensionalität stützt. Somit stehen die Ergebnisse in Einklang mit der Annahme über die Struktur des zugrundeliegenden Konstrukts und erfüllen damit dieses Kriterium für Validität nach Messick (1995).

Generalisierbarkeit

Die Generalisierbarkeit eines Tests lässt sich sowohl auf Item-Ebene als auch auf Personen-Ebene bestimmen und stellt ebenfalls einen Aspekt von Konstruktvalidität nach Messick (1995) dar. Auf der Item-Ebene wird betrachtet, inwiefern die Testergebnisse auf andere Items, die das gleiche Konstrukt messen, generalisiert werden können. Auf der Personen-Ebene wird betrachtet, inwie-

Tabelle 3. Interkorrelationen (Pearsons *r*) der Personenvariablen für konvergente und divergente Validität

	Interkorrelationen				
	1	2	3	4	5
1 WOR-TE	[.90]	.51 (.14)	.40 (.17)	.37 (.21)	.21 (.12)
2 Multiple Choice-Wortschatz (KFT)	.64 (.14)	[.64]	.28	.43	.22
3 Leseflüssigkeit (SLS)	.41 (.17)	.36	[.96]	.49	.04
4 Schreibfähigkeit (HSP)	.38 (.21)	.54	.50	[.99]	.19
5 Nonverbale Intelligenz (CFT)	.26 (.14)	.34	.05	.23	[.67]

Anmerkungen: Die obere Dreiecksmatrix enthält die manifesten, die untere die minderungskorrigierten Korrelationen. Die Reliabilität ist in der Diagonale in eckigen Klammern angegeben. Da die standardisierten Instrumente altersspezifische Werte ergeben, wurden die Korrelationen mit dem WOR-TE für jede Klassenstufe separat berechnet und anschließend gemittelt. Standardabweichungen sind in runden Klammern angegeben.

fern die Testergebnisse auf andere Populationen generalisiert werden können. Eine Möglichkeit, Generalisierbarkeit zu messen, ist die Kreuzvalidierung von Ergebnissen mithilfe verschiedener Teilungskriterien. Auf der Item-Ebene wurden die 204 Items in zwei Gruppen geteilt und je ein neues Modell pro Gruppe geschätzt, wie bereits zur Prüfung der Modellpassung angegeben. Anschließend wurden die Personenparameter beider Modelle verglichen. Mit einer Korrelation von $r = .94$ zeigt sich ein starker Zusammenhang, was darauf hinweist, dass die Personenparameter auch mit verschiedenen Item-Gruppen hergestellt werden können. Auf der Personen-Ebene wurden die teilnehmenden Kinder in zwei Gruppen geteilt und jeweils ein neues Modell für jede Gruppe berechnet. Anschließend wurde die Korrelation der Itemparameter beider Modelle berechnet. Mit $r = .97$ kann diese als sehr hoch bewertet werden. Mit verschiedenen Stichproben werden demnach sehr ähnliche Itemparameter gemessen.

Die Testergebnisse sind demnach durchaus generalisierbar, sowohl auf der Item- als auch auf der Personen-Ebene, was wiederum die Validität des Tests laut Messicks Definition (1995) unterstreicht.

Externe Aspekte

Der Zusammenhang zwischen Testergebnissen und anderen externen Variablen ist ein weiterer Aspekt von Validität. Messick (1995) schlägt vor, dabei sowohl konvergente Variablen, die eng mit dem zu messenden Konstrukt zusammenhängen, als auch divergente Konstrukte, die nur schwach oder gar nicht mit den Testergebnissen in Verbindung stehen, in Betracht zu ziehen.

In vorherigen Studien ist der starke Zusammenhang zwischen mündlichen Definitionen und den Ergebnissen aus Ja / Nein-Wortschatztests häufig gezeigt worden (z. B. Anderson & Freebody, 1983; Mochida & Harrington, 2006). Für die vorliegende Untersuchung wurden in einer Pilotstudie Daten zur mündlichen Definition von Kindern erhoben. Die teilnehmenden Kinder ($N = 27$, Alter $M = 10.3$, $SD = 0.57$) wurden nach Durchführung des Wortschatztests WOR-TE aufgefordert, mündliche Definitionen, sowohl zu einem Teil der angekreuzten als auch zu einem Teil der nicht angekreuzten Wörter, zu geben. Die Definitionen wurden auf Grundlage ihres semantischen Gehalts auf einer Skala von 0 bis 3 Punkten in Anlehnung an Gutierrez-Cleflen und DeCurtis (1999) bewertet. Die Ergebnisse zeigten zum einen, dass angekreuzte Wörter besser definiert werden konnten ($M = 1.07$, $SD = 0.93$) als nicht angekreuzte ($M = 0.30$, $SD = 0.68$). Zudem erwies sich ein hoher Zusammenhang zwischen dem summierten Definitionsergebnis und dem Personenparameter im WOR-TE, $r = .69$. Dies ist vergleichbar mit vorherigen Studien und weist darauf hin, dass der mit dem WOR-TE gemessene orthographische Wortschatz auch eng mit semantischen Kenntnissen über Wörter verbunden ist.

Zusätzlich enthielt die vorliegende Studie zur Messung der konvergenten Validität mehrere (standardisierte) Instrumente, die Variablen, die eng mit dem orthographischen Wortschatz verknüpft sind, erheben. Dazu wurden der Wortschatz mit einem Multiple-Choice-Verfahren (KFT) gemessen, die Leseflüssigkeit (SLS) und die Schreibfähigkeit (HSP) erhoben.

Tabelle 3 zeigt die Interkorrelationen der Personenvariablen (manifeste im oberen, minderungskorrigierte im unteren Dreieck). Es konnten wie erwartet moderate bis hohe Korrelationen des Personenparameters aus dem Ja / Nein-Wortschatztest mit den anderen Konstrukten gemessen werden. Ein größerer orthographischer Wortschatz hängt somit eng mit dem Wortschatz, der Leseflüssigkeit und der Schreibfähigkeit zusammen. Dies steht in Einklang mit vorherigen Studien zum Zusammenhang des Wortschatzes zu anderen Variablen (Anderson, Wilson, & Fielding, 1988; Aarnoutse, van Leeuwe, Voeten, & Oud, 2001) und zeigt, dass der WOR-TE tatsächlich den Wortschatz erfasst.

Zur Messung der divergenten Validität wurde die nonverbale Intelligenz mithilfe eines CFT-Subtests (Matrizen) erhoben. Die Korrelation mit dem WOR-TE ist ebenfalls in Tabelle 3 dargestellt. Wie erwartet fällt sie relativ gering aus, $r = .26$. Ähnliche Ergebnisse wurden in anderen deutschen Wortschatztests gefunden (z. B. WWT, Glück, 2011).

Bezüglich der externen Aspekte der konvergenten Validität konnten plausible Korrelationen für den WOR-TE mit anderen, dem orthographischen Wortschatz nahen Konstrukten gefunden werden. Zudem zeigen Daten aus einer Pilotstudie mit mündlichen Definitionen ähnliche Ergebnisse wie frühere Studien zu Ja / Nein-Wortschatztests. Für die divergente Validität wurde ein geringer Zusammenhang zwischen Wortschatz und nonverbaler Intelligenz gezeigt. Die Ergebnisse zur externen Validität sind damit zufriedenstellend.

Gültigkeit des Rasch-Modells

Ein wesentlicher Aspekt bei der Verwendung des Rasch-Modells für eine Testanalyse ist die Prüfung der Gültigkeit des Modells. Zwar liegt dafür kein allgemeingültiges Verfahren vor, dennoch können verschiedene Analysen, die die Annahmen des Modells bestätigen, zur Prüfung der Passung herangezogen werden (Rost, 1999). Viele dieser Analysen sind bereits im vorgestellten Validitätskonzept enthalten.

Zum einen betrifft dies Analysen, die sich auf die Passung des Modells auf den Datensatz beziehen. In der vorliegenden Analyse sind dazu die Split-Half-Korrelationen heranzuziehen. Sowohl eine Aufteilung der Items in zwei Gruppen als auch eine Aufteilung der Personen in zwei Gruppen ergab eine hohe Korrelation der jeweiligen korrespondierenden Parameter. Die Modellannahme der Stichprobenunabhängigkeit ist damit bestätigt und spricht für die Modellpassung. Die Modellannahme der lokalen stochastischen Unabhängigkeit konnte zudem mit der Q3-Statistik unterstrichen werden. Zum anderen können zur Prüfung des Modells Vergleiche mit anderen Modellen, die aus theoretischer Sicht sinnvoll sind und ebenfalls auf die Daten passen könnten, in Betracht gezogen werden (Rost, 1999). Hierzu wurde das vorliegende Modell mit Modellen mit mehreren Dimensionen verglichen, zum einen auf Ebene der Testversionen, zum anderen auf Ebene der Wortarten. Aus theoretischer Sicht liegt darin die Annahme, dass in den verschiedenen Altersgruppen (Testversionen) unterschiedliche Fähigkeiten zur Lösung des Tests benötigt werden bzw. für die verschiedenen Wortarten jeweils andere Kompetenzen gefragt sind. Beide Modelle zeigten keinen signifikanten Unterschied zum ursprünglichen Modell, was dessen Passung ebenfalls unterstreicht. Zusätzlich wurde zur Überprüfung der Modell-Passung der Likelihood-Ratio-Test nach Andersen (Glas & Verhelst, 1995) einzeln für ein Modell pro Klassenstufe geschätzt. Lediglich für die 2. Klasse ergab sich ein leicht signifikantes Ergebnis ($p = .04$), in allen anderen Klassenstufen war der Test nicht signifikant (alle $p > .1$). Zusammenfassend kann davon ausgegangen werden, dass das Modell ausreichend auf die Daten passt.

Diskussion

In diesem Artikel wurde der Ja / Nein-Wortschatztest WOR-TE für Grundschulkinder vorgestellt und anhand des Rasch-Modells validiert. Der Test enthält drei Versionen für verschiedene Altersgruppen und kann im Gruppensetting innerhalb von kurzer Zeit angewendet werden. Gegenüber anderen Verfahren (z. B. PPVT-4; Lenhard, Lenhard, Segerer & Suggate, 2015; PDSS, Kauschke & Siegmüller, 2009) hat er damit den klaren Vorteil, dass er in einer Gruppensituation mit mehreren Kindern angewendet werden kann. Zudem ist er für eine Altersgruppe konzipiert, für deren Messung im Bereich Wortschatz bisher wenige Verfahren vorlagen. Bereits existierende Verfahren (z. B. WWT; Glück, 2011) zielen eher auf die Diagnostik semantisch-lexikalischer Defizite ab. Der WOR-TE hingegen ist eher ressourcenorientiert und zur Messung des orthographischen Wortschatzes von Kindern geeignet. Gegenüber anderen Verfahren, die beispielsweise das Finden von Synonymen beinhalten (z. B. CFT-20-R; Weiß, 2006) hat der WOR-TE den Vorteil, dass die Abhängigkeit von den Distraktor-Items, in diesem Fall die Pseudowörter, relativ gering ist, was sich in den geringen Korrelationen mit dem Testverhalten gezeigt hat. Allen anderen Testverfahren hat der WOR-TE zudem die hohe Anzahl an Test-Items, die durch das einfache Testformat begründet sind, voraus.

In der vorliegenden Studie wurde versucht, anhand des Validitätskonzepts nach Messick (1995) Evidenz für die Validität des WOR-TE zu finden. Diese Aspekte umfassen inhaltliche, substanzielle, strukturelle und externe Aspekte sowie Generalisierbarkeit. Die Gültigkeit des Rasch-Modells wurde anhand verschiedener Aspekte als ausreichend betrachtet.

Zusammengefasst liegen starke Hinweise für die Validität des Verfahrens zur Messung des kindlichen orthographischen Wortschatzes vor. Es wurden zufriedenstellende Ergebnisse für alle von Messick (1995) vorgeschlagenen Aspekte der Validität erzielt. Die Analysen gaben zudem Hinweise auf Möglichkeiten zur weiteren Verbesserung des Verfahrens, insbesondere bezüglich der Anzahl der Items und der Itemauswahl. Alles in allem be-

steht eine starke Evidenz dafür, dass es sich bei dem WOR-TE um ein valides Instrument zur Erfassung des orthographischen Wortschatzes bei Grundschulkindern im Deutschen handelt. Insbesondere in forschungsbezogenen Kontexten stellt er damit eine gute Option zur Erfassung des kindlichen Wortschatzes dar. Da keine Normwerte vorliegen, ist eine Individualdiagnose derzeit mit dem Instrument jedoch nicht möglich.

Während die Auswertung von Ja / Nein-Wortschatztests in vorherigen Studien häufig anhand der Hits und False-Alarm-Raten erfolgte (z. B. Eyckmans, 2004; Huibregtse, Admiraal, & Merea, 2002), wurde in der vorliegenden Studie lediglich auf die Hits zurückgegriffen, um die Auswertung mit dem Rasch-Modell zu ermöglichen. Mochida und Harrington (2006) konnten bereits zeigen, dass die alleinige Auswertung der Hits am besten mit anderen Wortschatzmaßen korrelierte. Auch unsere Ergebnisse sprechen dafür, dass die Korrektur mithilfe der False-Alarm-Rate nicht notwendig ist. Es bestand nur eine schwache Beziehung zwischen Rateverhalten und Personenparameter, die sogar tendenziell darauf hinwies, dass Kinder mit höherer Ratetendenz einen geringeren Personenscore hatten und damit weniger Wörter angekreuzt haben. Korrigiert man anhand der False-Alarm-Rate, geht man davon aus, dass Personen, die mehr raten, auch generell *zu viele* Wörter angekreuzt haben (Mochida & Harrington, 2006), was in den vorliegenden Daten nicht der Fall ist. Dies rechtfertigt zunächst die alleinige Verwendung der Hit-Raten. Dennoch werden weiterhin verschiedene Methoden der Auswertung und Korrektur von Ergebnissen von Ja / Nein-Wortschatztests diskutiert (vgl. Huibregtse, Admiraal, & Merea, 2002; Pellicer-Sánchez & Schmitt, 2012). Eine weitere Analyse der vorliegenden Daten könnte hier weitergehende Erkenntnisse liefern.

Einschränkend lässt sich zudem festhalten, dass der Einsatz des Tests bei Leseanfängerinnen und Leseanfängern kritisch zu sehen ist. Dies zeigt sich durch die erhöhten Ratetendenzen in der 1. und 2. Klasse. Die Lesefähigkeit ist in diesen Klassenstufen wohlmöglich noch zu gering, sodass eine Erhebung des orthographischen Wortschatzes erst später möglich ist. Limitierend ist für diese junge Altersgruppe auch die Reliabilität im standardisierten Wortschatztest aus dem KFT zu nennen. Möglicherweise ist der Wortschatz in diesem Alter von Kind zu Kind sehr unterschiedlich (siehe auch Segbers & Schroeder, 2017), was eine reliable Messung mithilfe einer kleinen Item-Anzahl erschwert. Ein Vergleich der Testdaten mit auditiv vorgegebenen Wörtern in höherer Anzahl als Ja / Nein-Verfahren könnte hier eine sinnvolle Ergänzung sein.

Die Analyse des Effekts von Mehrsprachigkeit auf das Testergebnis zeigte, dass es Unterschiede zwischen ein- und mehrsprachigen Kindern im Testverhalten gibt. Eine detailliertere Analyse dieser Unterschiede könnte Aufschluss darüber geben, inwiefern der Einsatz des Tests bei mehrsprachigen Kindern sinnvoll ist bzw. die Ergebnisse mit denen der einsprachigen Kinder vergleichbar sind. Weiterhin lässt sich anmerken, dass das Verfahren nicht zur Erfassung von detailliertem Wortschatzwissen, insbesondere auf der semantischen Ebene, geeignet ist. Zwar sind die Ergebnisse aus der Pilotstudie mit den mündlichen Definitionsaufgaben vielversprechend, dennoch können mit dem WOR-TE keine detaillierten Aussagen über das semantische Wissen gemacht werden. Um dieses zu erfassen und eine differenzierte Individualdiagnose zu erstellen, sind aufwendigere Testverfahren von Nöten. Bei dem vorgestellten Instrument handelt es sich also um eine Möglichkeit zur Erfassung des orthographischen Wortschatzes, der substanziell mit dem semantischen Wortschatz zusammenhängt.

Zusätzlich konnte gezeigt werden, dass die Verwendung des Validitätskonzepts nach Messick (1995) die Möglichkeiten zur Validierung eines Ja / Nein-Wortschatztests über die üblichen Korrelationen mit mündlichen Definitionen oder Multiple-Choice-Fragen hinaus erweitert. Die vorliegenden Analysen beinhalteten relevante Schritte zur Sicherung von Evidenz für die Validität eines Verfahrens und zur Absicherung und Verbesserung der Qualität eines Instruments. Das vorgestellte Vorgehen zur Validierung kann damit als wichtiger Beitrag für die Entwicklung von Ja / Nein-Wortschatztests angesehen werden und sollte für die zukünftige Konstruktion ähnlicher Instrumente in Betracht gezogen werden.

Elektronische Supplemente (ESM)

Die elektronischen Supplemente sind mit der Online-Version dieses Artikels verfügbar unter https://doi.org/10.1026/0012-1924/a000212

ESM 1. Items

Literatur

Aarnoutse, C., van Leeuwe, J., Voeten, M. & Oud, H. (2001). Development of decoding, reading comprehension, vocabulary and spelling during the elementary school years. *Reading and Writing: An Interdisiplinary Journal*, *14*, 61 – 89. https://doi.org/10.1023/A:1008128417862

Anderson, R. C. & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. *Advances in Reading/Language Research*, *2*, 231 – 256.

Anderson, R. C., Wilson, P. T. & Fielding, L. G. (1988). Growth in reading and how children spend their time outside of school. *Reading Research Quarterly*, *23*, 285 – 303.

Anglin, J. M., Miller, G. A. & Wakefield, P. C. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development, 58*, 1 – 186. https://doi.org/10.2307/1166112

Auer, M., Gruber, G., Mayringer, H. & Wimmer, H. (2005). *Salzburger Lese-Screening für die Klassenstufe 5 – 8.* Bern: Hans Huber.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*, 101 – 118. https://doi.org/10.1177/0265532209340194

Bialystok, E., Luk, G., Peets, K. F. & Yang, S. (2010). Receptive vocabulary differences in monolingual and bilingual children. *Bilingualism: Language and Cognition*, *13*, 525 – 531. https://doi.org/10.1017/S1366728909990423

Biemiller, A. (2003). Vocabulary: Needed if more children are to read well. *Reading Psychology*, *24*, 232 – 335. https://doi.org/10.1080/02702710390227297

Biemiller, A. (2005). Size and sequence in vocabulary development: Implications for choosing words for primary grade vocabulary instruction. In A. Hiebert & M. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (pp. 223 – 242). Mahwah, NJ: Erlbaum.

Biemiller, A. (2006). Vocabulary development and instruction: A prerequisite for school learning. In D. K. Dickinson & S. B. Neuman (Eds.), *Handbook of Early Literacy Research* (Vol. 2, pp. 41 – 51). New York: Guilford Press.

Bock, R. D. & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 433 – 448). New York: Springer.

Cattell, R. B., Weiß, R. H. & Osterland, J. (1997). *Grundintelligenztest Skala 1.* Göttingen: Hogrefe.

Coltheart, M., Rastle, K., Perry, C., Langdon, R. & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204 – 256.

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists.* New Jersey: Lawrence Erlbaum.

Eyckmans, J. (2004). *Measuring receptive vocabulary size: Reliability and validity of the Yes/No Vocabulary Test for french-speaking learners of dutch.* Utrecht: LOT.

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., et al. (2006). The cognitive correlates of third grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology*, *98*, 29 – 43. https://doi.org/0.1037/0022-0663.98.1.29 – *Verweis fehlt im Text – bitte prüfen*

Glas, C. A. W. & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models* (pp. 69 – 95). New York: Springer.

Glück, C. W. (2011). *Wortschatz- und Wortfindungstest für 6- bis 10-Jährige.* Amsterdam: Elsevier.

Grimm, H. & Doil, H. (2006). *Elternfragebögen für die Früherkennung von Risikokindern* (ELFRA) (2. Aufl.). Göttingen: Hogrefe.

Gutierrez-Cleflen, V. F. & DeCurtis, L. (1999). Word definition skills in Spanish-speaking children with language impairment. *Communication Disorders Quarterly*, *21*, 23 – 31. https://doi.org/10.1177/152574019902100104

Heller, K. & Geisler, H. J. (1983). *Kognitiver Fähigkeitstest für 1. bis 3. Klassen.* Weinheim: Beltz.

Heller, K. A. & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen* (Revision). Göttingen: Beltz Test.

Huibregtse, I., Admiraal, W. & Merea, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing*, *19*, 227 – 245. https://doi.org/10.1191/0265532202lt229oa

Kauschke, C. & Siegmüller, J. (2009). *Patholinguistische Diagnostik bei Sprachentwicklungsstörungen* (2. Aufl.). Amsterdam: Elsevier.

Kiefer, T., Robitzsch, A. & Wu, M. (2016). *TAM: Test Analysis Modules* [Computer Software].

Kolen, M. J. & Brennan, R. L. (2004). *Test Equating, Scaling and Linking. Methods and Practices.* New York: Springer.

Lemhöfer, K. & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*, 325 – 343. https://doi.org/10.3758/s13428-011-0146-0

Lenhard, A., Lenhard, W., Segerer, R. & Suggate, S. (2015). *Peabody Picture Vocabulary Test* (4. Ausgabe: Deutsche Fassung). Frankfurt am Main: Pearson Assessment.

May, P. (2002). *Hamburger Schreibprobe 1 – 10.* Stuttgart: Ernst Klett Verlag.

Mayringer, H. & Wimmer, H. (2003). *Salzburger Lese-Screening für die Klassenstufe 1 – 4.* Bern: Hans Huber.

McLean, S., Kramer, B. & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*, 741 – 760. https://doi.org/10.1177/1362168814567889

McNamara, T. (1996). *Measuring second language performance.* Harlow: Addison Wesley Longman.

Merea, P. & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*, 142 – 154.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741 – 749.

Mislevy, R. J. & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, *13*, 57 – 75.

Mochida, K. & Harrington, M. (2006). The yes/no test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*, 73 – 98. https://doi.org/10.1191/0265532206lt321oa

Muter, V., Hulme, C., Snowling, M. J. & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: Evidence from a longitudinal study. *Developmental Psychology*, *40*, 665 – 681. https://doi.org/10.1037/0012-1649.40.5.665

Ouellette, G. & Beers, A. (2010). A not-so-simple view of reading: How oral vocabulary and visual-word recognition complicate the story. *Reading and Writing*, *23*, 189 – 208. https://doi.org/10.1007/s11145-008-9159-1

Pearson, P. D., Hiebert, E. H. & Kamil, M. L. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly*, *42*, 282 – 298. https://doi.org/10.1598/RRQ.42.2.4

Pellicer-Sánchez, A. & Schmitt, N. (2012). Scoring yes-no vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29*, 489 – 509. https://doi.org/10.1177/0265532212438053

Perfetti, C. A. & Hart, L. (2002). The lexical quality hypothesis. *Precursors of Functional Literacy*, *11*, 67 – 86.

Perfetti, C. & Stafura, J. (2013). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*, 22 – 37. https://doi.org/10.1080/10888438.2013.827687

Purpura, D. J., Hume, L. E., Sims, D. M. & Lonigan, C. J. (2011). Early literacy and early numeracy: The value of including early literacy skills in the prediction of numeracy development. *Journal of Experimental Child Psychology, 110*, 641 – 658. https://doi.org/10.1016/j.jecp.2011.07.004

Ricketts, J., Nation, K. & Bishop, D. V. M. (2007). Vocabulary is important for some, but not all reading skills. *Scientific Studies of Reading*, *11*, 235 – 257. https://doi.org/10.1080/10888430701344306

Rost, J. (1999). Was ist aus dem Rasch-Modell geworden? *Psychologische Rundschau, 50*, 140 – 156. https://doi.org/10.1026//0033 – 3042.50.3.140

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A. & Kliegl, R. (2015). childLex: A lexical database of German read by children. *Behavior Research Methods*, *47*, 1085 – 1094. https://doi.org/10.3758/s13428-014-0528-1

Schröter, P. & Schroeder, S. (2017). The developmental lexicon project: A behavioral database to investigate visual word recognition across the lifespan. *Behavior Research Methods, 47*, 2183 – 2203. https://doi.org/10.3758/s13428-016-0851-9

Segbers, J. & Schroeder, S. (2017). How many words do children know? A corpus-based estimation of children's total vocabulary size. *Language Testing, 34*, 297 – 320. https://doi.org/10.1177/0265532216641152

Shillaw, J. (1996). The application of Rasch modelling to yes/no vocabulary tests. *Vocabulary Acquisition Research Group*, University of Wales: Swansea.

Weiß, R. H. (2006). *Grundintelligenztest Skala 2 – Revision mit Wortschatztest und Zahlenfolgentest* (Revision). Göttingen: Hogrefe.

Welsh, J. A., Nix, R. L., Blair, C., Bierman, K. L. & Nelson, K. E. (2010). The development of cognitive skills and gains in academic school readiness for children from low-income families. *Journal of Educational Psychology, 102*, 43 – 53. https://doi.org/10.1037/a0016738

Yen, W. M. (1984). Effect of local item dependence on fit and equating performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement, 8*, 125 – 145.

Onlineveröffentlichung: 25. 09. 2018

Dr. Sascha Schroeder
Jutta Trautwein
Max-Planck-Institut für Bildungsforschung
MPFG Reading Education and Development (REaD)
Lentzeallee 94
14185 Berlin
sascha.schroeder@mpib-berlin.mpg.de

