Empirische Sonderpädagogik, 2017, Nr. 3, S. 236-257
ISSN 1869-4845 (Print) · ISSN 1869-4934 (Internet)

Konzeption und Güte curriculumbasierter Messverfahren zur Erfassung der arithmetischen Leistungsentwicklung in den Klassenstufen 3 und 4

Simon Sikora & Stefan Voß

Universität Rostock

Zusammenfassung

Die formative Leistungsdiagnostik erfährt international wie auch national ein hohes Forschungsinteresse. Insbesondere der US-amerikanische Ansatz der Curriculum-based Measurements (CBM; Deno, 1985, 2003; Fuchs, 2004) erscheint vielversprechend, wenn es um die kurzfristige Analyse der Leistungsentwicklung in der Schule geht. Vor diesem Hintergrund wird im vorliegenden Beitrag über die Vorgehensweise bei der Konstruktion und Evaluation von CBM für jede der vier Grundrechenarten berichtet, welche im Abstand von vier Schulwochen eingesetzt werden können. In einem ersten Schritt wurden hierzu nach mathematikdidaktischen sowie empirischen Gesichtspunkten geclusterte Itempools generiert, die jeweils Grundlage zur Extraktion stratifizierter, d. h. strukturähnlicher CBM hinsichtlich Addition, Subtraktion, Multiplikation und Division für den Mathematikunterricht der 3. und 4. Klasse darstellten.
Mittels eines Multi-Matrix-Designs konnten in einem zweiten Schritt empirische Kennwerte zur Güte der entwickelten CBM in einer Stichprobe von N = 463 Grundschülerinnen und Grundschülern klassenstufenübergreifend zu sechs Messzeitpunkten über zwei Schuljahre erhoben werden. In Teilstichproben wurden die CBM in vierwöchentlichem Abstand durchgeführt sowie weitere Instrumente zur Validierung der entwickelten Verfahren eingesetzt. Die Datenlage erlaubt differenzierte Aussagen hinsichtlich der statusdiagnostischen (Reliabilität und Validität) und darüber hinaus der verlaufsdiagnostischen Eignung (Skalierung nach IRT und mehrebenenanalytische Modellierung der Änderungssensibilität) der entwickelten CBM.
Zusammenfassend ist festzuhalten, dass es grundlegend gelungen ist, schwierigkeitsähnliche, reliable wie auch valide CBM-Versionen für jede Grundrechenoperation zu generieren. Einschränkungen werden diskutiert und weitere Desiderata in Ausblick gestellt.

Schlagwörter: Lernverlaufsdiagnostik, Arithmetik, Curriculumbasierte Messverfahren, Item-Response-Theorie, Grundschule

Conception and quality of curriculum-based measurments for the computation performance of primary school students in grade 3 and 4

Abstract

The progress monitoring of school performace is in the spotlight of international and national researchers. In particular, the US-American curriculum-based measurements approach (CBM; Deno, 1985, 2003, Fuchs, 2004) seems promising when it comes to the short-term analysis of

students' performance in school. The present paper deals with the construction and evaluation process of computation CBM for each of the four basic operations, which can be used in four-week intervals. In a first step, clustered pools of items were generated according to mathematical and empirical aspects. Then structural similar CBM for primary school students in grades 3 and 4 were extracted on the basis of the item clusters.
In a second step, empirical values concerning the quality of the developed CBM were assessed in a sample of N = 463 primary-school students across grades 3 and 4 with a multi-matrix-design. In subsamples, the measurements were used in four-week intervals as well as further instruments to validate the developed CBM. The data presented allows differentiated statements regarding the suitability to measure status (reliability and validity) and, in addition, progress (scaling according to IRT and multi-level modeling of change sensitivity) of the tested children. In summary, it can be said that it has been fundamentally successful to generate reliable and valid CBM forms with similar difficulty for each basic operation. Restrictions are discussed and further desiderata are presented.

Keywords: progress monitoring, arithmetics, curriculum-based measurement, Item Response Theory, elementary school

Die formative Evaluation gilt als ein wirksames Element zur Planung und Überprüfung von Unterrichts- und Fördermaßnahmen (z. B. Hattie, 2013). Während in den USA bereits seit den 1970er Jahren intensiv an sog. curriculum-based measurements (CBM; u. a. Deno, 1985, 2003; Fuchs, 2004) als eine Methode der formativen Diagnostik geforscht wird, mangelt es im deutschsprachigen Raum derzeit noch an Testverfahren, welche in der Lage sind, Leistungsentwicklungen zuverlässig wie auch sensibel über die Zeit zu erfassen (Voß & Hartke, 2014; Voß, Sikora & Hartke, 2017). Bereits vorhandene CBM für den Mathematikunterricht im Grundschulalter sind zudem nicht bzw. nicht hinreichend normiert, sodass die Leistungsentwicklung vorrangig im intraindividuellen Vergleich betrachtet werden muss (Strathmann & Klauer, 2012; Hartmann & Müller, 2014).

Inhaltliche Abwägungen bei der Konzeption von Instrumenten zur Lernverlaufsdiagnostik ebenso wie *methodische Herausforderungen* bei deren Evaluation stellen einen hohen Anspruch an die Forschung. Inhaltlich sind in erster Linie Fragen nach dem zu messenden Konstrukt zu klären, wie es auch bei der Konstruktion von statusdiagnostischen Verfahren üblich ist. Diese Überlegungen müssen erweitert werden um Gedanken zur Erfassung von Entwicklungen im zu messenden Konstrukt. Beispielhafte Fragen könnten sein:

– Gibt es Theorien oder Modelle, welche die Entwicklung im zu prüfenden Konstrukt beschreiben und somit eine Orientierung bei der Konzeption von Testverfahren liefern können? Sind Entwicklungsstufen auszumachen, sodass eine Verortung des Kindes in diesem Modell ermöglicht wird? Wie können diese Stufen operationalisiert werden?
– Handelt es sich um ein ein- oder mehrdimensionales Konstrukt? Sind überhaupt kurzfristige Änderungen zu erwarten und wenn ja, in welchem Ausmaß?

Analog dazu ist aus methodischer Sicht sowohl gängigen statusdiagnostischen (psychometrische Güte des Verfahrens hinsichtlich der Testgütekriterien Objektivität, Reliabilität und Validität) sowie verlaufsbezogenen Anforderungen zu genügen (Fuchs, 2004; Wilbert & Linnemann, 2011; Voß et al., 2017). Hierbei geht es insbesondere um die Frage nach einem geeigneten Messmodell zur Analyse der Eignung zur Veränderungsmessung. In diesem Zusammenhang besteht Einigkeit darüber, dass die klassische Testtheorie mit Lernverlaufsdiagnostik eher nicht vereinbar ist und es einer Ergänzung um ein entsprechendes probabilistisches Messmodell bedarf (Rost, 2004; Klauer, 2011; Wilbert & Linnemann, 2011).

Vor diesem Hintergrund wurden im Rahmen der hier beschriebenen Studie CBM für den Inhaltsbereich Arithmetik entwickelt, welche im vierwöchentlichen Abstand im Verlauf der dritten und vierten Klassenstufe eingesetzt werden können. Neben der Beschreibung der Konstruktion dieser Verfahren liegt der Fokus des Beitrags in der Analyse ihrer statusdiagnostischen sowie verlaufsdiagnostischen Eignung.

Inhaltliche Überlegungen zur Konstruktion der CBM

Zur Entwicklung von CBM, mit denen die mathematischen Lernfortschritte von Schülerinnen und Schülern am Ende der Grundschulzeit abgebildet werden können, muss zunächst geklärt werden, welche Ziele der Mathematikunterricht verfolgt. Es geht somit um die Frage nach dem zu erfassenden Konstrukt. Beim Blick in die mathematikdidaktische Fachliteratur (u. a. Grassmann, Eichler, Mirwald & Nitsch, 2014; Käpnick, 2014; Krauthausen & Scherer, 2007; Schipper, 2009) sowie in die Bildungsstandards für das Fach Mathematik im Primarbereich (KMK, 2005) wird deutlich, dass sich der Mathematikunterricht nicht in Zahlen und Operationen mit ihnen erschöpft. Neben der Arithmetik werden auch das Sachrechnen, der Umgang mit Größen, die Geometrie sowie die Stochastik thematisiert. Winkelmann, Robitzsch, Stanat und Köller (2012) zeigten durch eine methodisch überzeugende Studie mit etwa 17 000 Dritt- und Viertklässlern, dass diese Inhaltsbereiche statistisch diskriminierbar sind, mathematische Kompetenz somit ein mehrdimensionales Konstrukt ist. „Die These, nach der mathematische Kompetenz in all ihren Facetten bestmöglich durch einen globalen Faktor (g) erklärbar sei, wurde hier für die Primarstufe nicht bestätigt“ (Winkelmann et al., 2012, S. 24). Eine notwendige Voraussetzung zur Messung von Lernverläufen stellt jedoch die Eindimensionalität der Testskalen dar (Wilbert & Linnemann, 2011; Wilbert, 2014). Somit müssten Verfahren konstruiert werden, welche die Entwicklung in jeweils einer der inhaltlichen Dimensionen erfassen. Bisher veröffentlichte CBM fokussieren diesbezüglich auf den Inhaltsbereich der Arithmetik (Hartmann & Müller, 2014; Strathmann & Klauer, 2012). Dies erscheint logisch, da arithmetische Kompetenzen eine essenzielle Grundlage für ein sicheres Operieren in den weiteren Inhaltsbereichen der Grundschulmathematik darstellen. Ein profunder Umgang mit Größen, geometrischen Objekten, Daten und Wahrscheinlichkeiten setzt Zahlvorstellungen und Rechenfähigkeiten voraus. Die Arithmetik ist somit von zentraler Bedeutung für die mathematische Entwicklung.

Mit Blick auf die Prävention von Lernschwierigkeiten ist insbesondere der arithmetische Teilaspekt von Relevanz, da sich ausschließlich hierauf bei der formalen Diagnose einer Rechenstörung bzw. Dyskalkulie bezogen wird. Vor dem Hintergrund von Large-Scale Untersuchungen wie TIMSS ist anzunehmen, dass etwa ein Fünftel bis ein Viertel aller Schülerinnen und schüler im höheren Grundschulalter nicht über sichere Rechenfähigkeiten und -fertigkeiten verfügen (Selter, Walter, Walther & Wendt, 2016). Ähnlich hoch fällt der Anteil von Jugendlichen mit lediglich basalen mathematischen Kompetenzen im Sekundarstufenbereich aus (Gebhardt, Schwab, Schaupp, Rossmann & Gasteiger Klicpera, 2012; Hammer et al., 2015). Folglich ist davon auszugehen, dass eine Lernverlaufsdiagnostik im Bereich der Arithmetik für die Förderplanung hoch relevant ist.Für die vorliegende Studie ergibt sich daher das Ziel, das Konzept oben genannter, bereits verfügbarer CBM für den Mathematikunterricht aufzugreifen und deren Schwachstellen durch gezielte Modifikationen zu mildern. So sollten die entwickelten CBM ebenfalls die Kopfrechenfähigkeiten der Kinder in den vier Grundrechenarten prüfen.

Die erstellten CBM basieren auf dem Konzept des lehrziel- bzw. kriteriumsorientierten Testens (Klauer, 1987). Für jede der vier Grundrechenarten kann beschrieben werden, welche Anforderungen die Schülerinnen und Schüler am Ende der dritten bzw. vierten Klassenstufe erfüllen müssen. Einen Rahmen liefern die im jeweiligen Schuljahr behandelten Zahlenräume. Nach Krauthausen und Scherer (2007) rechnen Kinder in der dritten Klassenstufe im Tausenderraum, Kinder in der vierten Klassenstufe im Millionraum. So könnte von einem Drittklässler beispielsweise erwartet werden, dass er am Ende des Schuljahres in der Lage ist, alle erdenklichen Additionsaufgaben des Tausenderraums im Kopf zu lösen. In Abhängigkeit der bei der Rechnung zu verarbeitenden Stellen können verschiedene Schwierigkeitsgrade der Aufgaben unterschieden werden, wie Tabelle 1 exemplarisch verdeutlicht.

Vor der Erstellung der CBM wurden in einem ersten Schritt die möglichen Aufgabentypen in jeder Grundrechenart bestimmt (Definition der Aufgabenmenge) und nach Schwierigkeitsgrad geordnet. Diese Analyse umfasste auch den Zwanziger- sowie Hunderterraum, welche in den vorangehenden Klassenstufen thematisiert werden.

Aufgrund der großen Anzahl an Aufgabentypen mussten in einem zweiten Schritt Festlegungen getroffen werden, um repräsentative Teilmengen aus den Grundmengen in die CBM aufzunehmen (Klauer, 2011). Ein Itemsampling ohne entsprechende Restriktionen würde vermutlich nicht zu parallelen CBM führen. Grundgelegt wurden folgende vier Aspekte:

1. *Festlegung der zu berücksichtigenden Niveaustufen und der Anzahl der Aufgaben*

Für jede Grundrechenart wurde ein CBM entwickelt, das 24 Aufgaben umfasst. Jedes CBM besteht dabei aus 10 strukturgleichen Parallelversionen, die im Laufe eines Schuljahres im vierwöchentlichen Rhythmus eingesetzt werden können, ohne, dass gleiche Aufgabenblätter verwendet werden müssen. In den CBM sollten auch Anforderungen der zwei vorangegangenen Schuljahre enthalten sein, damit sie auch im unteren Leis-

Tabelle 1: Übersicht der Aufgabentypen bezüglich der Addition im Zahlenraum bis 1000

maximale Anzahl zu verrechnender Stellen	**Aufgabentyp**
1	Z + E, H + E, H + Z, H + H, HZ + E, HE + Z, H + ZE, HZ + H, HE + H, HZ + HE, HZE + H
2	E + E, Z + Z, HE + E, ZE + Z, HZ + Z, E + ZE, Z + ZE, HZ + ZE, HZ + HZ, HE + HE, HZE + Z, HZE + HZ
3	HE + ZE, HZE + E, HZE + ZE, HZE + HE, HZE + HZE

Anmerkungen: H – Hunderter; Z – Zehner, E – Einer

Tabelle 2: Verteilung der Anforderungen in den CBM

	Klasse 3	**Klasse 4**
Niveau Klasse 1 (ZR 20)	4 Aufgaben	
Niveau Klasse 2 (ZR 100)	12 Aufgaben	8 Aufgaben
Niveau Klasse 3 (ZR 1000)	8 Aufgaben	8 Aufgaben
Niveau Klasse 4 (ZR 1 000 000)		8 Aufgaben

Anmerkungen: ZR – Zahlenraum

tungsbereich gut differenzieren und somit auch in inklusiven Grundschulklassen sinnvollerweise eingesetzt werden können.

Der Schwierigkeitsgrad der Aufgaben eines CBM sollte durchgängig ansteigen. Dies wurde einerseits durch die stetige Erweiterung der Zahlenräume erreicht (vgl. Tabelle 2), zudem wurde die Anzahl der zu verarbeitenden Stellen (vgl. Tabelle 1) auch innerhalb einer Niveaustufe beständig erhöht. Damit nimmt die Zahl der zur jeweiligen Aufgabenlösung benötigten Rechenschritte sukzessive zu.

2. *Festlegung des maximalen Schwierigkeitsgrades*

Die CBM sollten ausschließlich Kopfrechenfähigkeiten abprüfen, da diese als ungemein wichtig anerkannt sind: „Auch im Zeitalter des Taschenrechners und Computers sowie des entdeckenden Lernens besteht in der Mathematikdidaktik Konsens über die schulische wie außerschulische Bedeutung solider Kopfrechenfertigkeiten und -fähigkeiten“ (Krauthausen & Scherer, 2007, S. 43). (Halb-)schriftliche Rechenverfahren wurden aus Gründen der Durchführungsökonomie nicht berücksichtigt, da deren Zeitaufwand dem CBM-Ansatz widersprechen würde.

Aufgrund dieses Beschlusses bedurfte es Überlegungen dahingehend, welche maximale Anforderung erwartet werden kann. Beispielsweise erscheint es nicht sinnvoll, Aufgaben wie 436 x 582 = _ zu stellen, da davon auszugehen ist, dass eine richtige Aufgabenlösung, aufgrund der Vielzahl an Rechenschritten, scheitern würde. Um den begrenzten Arbeitsgedächtniskapazitäten Rechnung zu tragen, wurde beschlossen, höchstens Aufgaben im Hunderttausenderraum (in Klasse 4) zu stellen. Zudem wurden nur Aufgabentypen ausgewählt, die maximal vier Rechenschritte erfordern. Bezüglich der Multiplikation und Division wurden nur Aufgaben genutzt, die mithilfe von Ableitungen aus Grundaufgaben lösbar sind.

3. *Festlegung der Aufgabenformen*

Die CBM sollten neben den formalen Kopfrechenaufgaben in der Standardform a + b = _ auch Gleichungen mit Platzhaltern nach den Formen a + _ = c bzw. _ + b = c umfassen. Deren Lösung setzt ein relationales Zahlverständnis (Fritz, Ricken & Gerlach, 2007) sowie ein Gleichungsverständnis voraus.

4. *Festlegung der Stellenüberschreitungen für jeden Aufgabentyp*

Zur Erzeugung paralleler Items war es bei vielen Aufgabentypen notwendig, die Anzahl der Stellenüberschreitungen zu standardisieren. Ohne eine entsprechende Festlegung würden beispielsweise die Aufgaben 111 + 111 = _ (kein Übergang) und 567 + 384 = _ (zwei Übergänge) demselben Aufgabentyp als Aufgaben ähnlicher Schwierigkeit zugeordnet werden.

Im Anschluss an die dargestellten Vorüberlegungen wurden die im jeweiligen CBM zu berücksichtigenden Aufgabentypen festgelegt sowie konkrete Items aus diesen Klassen per Zufallsprinzip ausgewählt. Bei diesem Vorgehen handelt es sich also, wie bereits bei Vorarbeiten mit ähnlicher Zielstellung (u. a. Strathmann & Klauer, 2010), um ein stratifiziert-zufälliges Verfahren der Aufgabenerzeugung, da sich die Grundmengen in jedem Zahlenraum aus klar unterscheidbaren Teilmengen mit unterschiedlichen Schwierigkeitsgraden zusammensetzen. Tabelle 3 stellt die ausgewählten Aufgaben im Bereich der Addition bezogen auf die Anforderungen der dritten Klassenstufe dar. Daraus geht hervor, dass nicht alle theoretisch möglichen Additionsaufgaben im Verfahren berücksichtigt werden konnten.

Von jedem ausgewählten Aufgabentyp wurden zehn verschiedene Kopfrechenaufgaben für jedes Schuljahr zufällig gezogen und auf zehn CBM-Versionen verteilt. Diesem Vorgehen liegt die Annahme zugrunde,

Tabelle 3: Anforderungen der dritten Klassenstufe im Bereich der Addition

Aufgabenstruktur	Übergänge	erster Summand größer	zweiter Summand größer	erster Summand größer	zweiter Summand größer	erster Summand größer	zweiter Summand größer	Beispiel
		a + b = _	a + b = _	a + _ = c	a + _ = c	_ + b = c	_ + b = c	
ZE + Z = HZE	1	X						87 + 30 = _
HZ + H = HZ	0		X					350 + 400 _
HZ + Z = HZ	1			X				750 + _ = 820
HZ + HZ = HZ	1					X		_ 370 = 930
HZE + HZ = HZ	1	X						457 + 360 = _
HZE + HZE = HZE	2		X					334 + 587 = _
HZ + HZ = HZ	1				X			250 + _ = 640
HZE + HZE = HZE	2						X	_ + 548 = 821

Anmerkungen: H – Hunderter; Z – Zehner, E – Einer

dass derart entwickelte CBM-Versionen parallel, d. h. gleich schwer sind und dasselbe messen. So ist beispielsweise davon auszugehen, dass es sich bei den Aufgaben 500 + 300 = _ und 500 + 400 = _ (Aufgabenstruktur HZ + H = HZ ohne Übergang; vgl. Tabelle 3) um parallele Items handelt. In einer ersten Pilotstudie zur Überprüfung der Parallelität der entwickelten Verfahren wurden beispielhaft je zwei CBM einer Grundrechenart in vier Klassen (N = 81) durchgeführt und die erreichten Rohwerte der Kinder miteinander korreliert. Dabei zeigten sich durchgängig hoch signifikante Korrelationskoeffizienten zwischen r = .72 und r = .86. Weiterführende Analysen der Güte der entwickelten CBM sind das Ziel der nachfolgend dargestellten Studie.

Da in den Verfahren der vierten Klassenstufe auch Aufgaben der dritten enthalten sind (Niveaustufen 2 und 3; vgl. Tabelle 2), liegt den erstellten CBM ein Multi-Matrix-Design (Mislevy, Beaton, Kaplan & Sheehan, 1992) zugrunde. Mittels Anker-Items wird, die Gültigkeit eines probabilistischen Messmodells vorausgesetzt, eine Modellierung des Lernverlaufs über beide Klassenstufen hinweg ermöglicht.

Zuletzt wurden Festlegungen zum Durchführungsmodus getroffen. Die Schülerinnen und Schüler sollen innerhalb von drei Minuten so viele Aufgaben einer Grundrechenart wie möglich im Kopf lösen. Die CBM sind somit als Speed-Tests konzipiert. Es ist nicht davon auszugehen, dass die Schülerinnen und Schüler alle Aufgaben in der vorgegebenen Zeit bearbeiten können. Die Konstruktion des Verfahrens folgte der Grundannahme, dass Kinder mit ineffektiven, umständlichen oder unsicheren Lösungsstrategien weniger Aufgaben in einer vorgegebenen Zeit lösen können als geschickte, kompetente und geübte Rechner. Diese Annahme wird durch entsprechende empirische Befunde gestützt (Voß, 2016). Durch die Zeitbegrenzung der CBM sollte folglich eine höhere Differenzierungsfähigkeit erreicht werden, bei gleichzeitiger Inkaufnahme eventueller Einbußen hinsichtlich der Änderungssensibilität.

Fragestellugen

Der vorliegende Beitrag soll nun über die psychometrische Güte der zuvor beschriebenen CBM informieren. Dabei wird folgenden forschungsleitenden Fragen nachgegangen:

F1 Wie fällt die psychometrische Güte der Items aus?
F2 Handelt es sich bei den konzipierten CBM um valide wie auch änderungssensible Verfahren?

Aus der Frage 1 resultieren folgende Hypothesen:

H1 Die Itemschwierigkeiten streuen über den gesamten Leistungsbereich.
H2 Die Trennschärfen betragen zumindest r_{pbis} = .20.
H3 Infit-Statistiken zum eindimensionalen Raschmodell liegen in einem Bereich zwischen $0.70 \leq \text{Infit} \leq 1.30$.
H4 Die einzelnen CBM sind hinreichend homogen ($\alpha > .80$).
H5 Korrelationen zwischen aufeinanderfolgenden Testzeitpunkten betragen zumindest r = .70 (Retest-Reliabilität im Abstand von 4 Wochen).

Hinsichtlich der Frage 2 ergeben sich weitere Hypothesen:

H6 Zwischen den CBM und konstruktnahen Außenkriterien gibt es einen hohen Zusammenhang ($r \geq .50$).
H7 Zwischen den CBM und konstruktfernen Außenkriterien gibt es einen geringeren Zusammenhang ($r < .50$).
H8 Die mittleren Zuwächse in den CBM-Rohwerten erlauben kurzfristige Aussagen zum Lernverlauf der Kinder.

Methode

Studienaufbau und Stichproben

Im Rahmen einer längsschnittlich angelegten Untersuchung wurden mittels der oben skizzierten CBM die mathematischen Leistungen von 463 Schülerinnen und Schülern der Region Rügen über die Schuljahre 2014/15 (Klasse 3) und 2015/16 (Klasse 4) jeweils zu drei Messzeitpunkten (4., 20. sowie 40. Schulwoche) erhoben. Die Kinder verteilten sich auf 20 Grundschulklassen in 11 Schulen. Das Geschlechterverhältnis war annähernd ausgeglichen (♂ = 234 / ♀ = 229), das mittlere Alter lag zum ersten Messzeitpunkt bei 8;9 Jahren (SD = 0;6 Jahre). Bei der Durchführung der CBM-Testungen wurde auf eine Begrenzung der Zeit, wie sie sonst üblich ist, verzichtet, um Kennwerte für jedes Item berechnen zu können.

Die Daten der Gesamtstichprobe wurden zur Beantwortung der Hypothesen 1-4 (psychometrische Güte der Items) genutzt. Um zusätzliche Informationen zur Güte der entwickelten CBM zu erhalten, wurden in zwei Teilstichproben weitere Daten erhoben, die einen Beitrag zur Klärung der Hypothesen 5-8 leisten können.

In acht der zuvor aufgeführten Klassen aus sieben Schulen wurden über die Schuljahre 2014/15 sowie 2015/16 zudem Testungen mit den oben beschriebenen CBM im monatlichen Abstand durchgeführt (8., 12., 16., 24., 28., 32. und 36. Schulwoche). Bei diesen wurde die Bearbeitungszeit auf drei Minuten pro CBM begrenzt. Wie in längsschnittlichen Studien üblich, zeigte sich über die Zeit ein Datenausfall in der untersuchten Teilgruppe. Es wurden die Werte all jener Kinder in die Analysen einbezogen, die an mindestens 45 der insgesamt 56 Testungen (7 Testungen je Schuljahr je Grundrechenoperation) teilgenommen hatten (80 %ige Datenverfügbarkeit). Diese Teilstichprobe 1 umfasste 152 Kinder (♂ = 71 / ♀ = 81), das mittlere Alter zum ersten Messzeitpunkt betrug 8;9 Jahre (SD = 0;5 Jahre). Das verwendete statistische Verfahren zur Modellierung der Lernverläufe (HLM, s. u.) ist jedoch sehr robust und erlaubt, im Gegensatz zu Strukturgleichungsmodellen, auch aussagekräftige Parameterschätzungen anhand relativ kleiner Stichproben (Bryk & Raudenbush, 1992).

In fünf der zuvor aufgeführten Klassen aus drei Schulen wurden zudem jeweils zum Beginn als auch zur Mitte der beiden Schuljahre die Mathematik- und die Deutschleistungen mit den entsprechenden KEKS-Tests (May & Bennöhr, 2013) als Außenkriterien erhoben. Diese Teilstichprobe 2 umfasste 69 Kinder (♂ = 31 / ♀ = 38), das mittlere Alter zum ersten Messzeitpunkt betrug 8;9 Jahre (SD = 0;5 Jahre). Alle mit dieser Untersuchung gekoppelten Testungen führten die Lehrkräfte eigenständig durch. Sie werteten die Verfahren ebenfalls selbst aus und trugen die Testergebnisse in eine Internetplattform ein (Voß et al., 2016), auf der sie jederzeit Einblick in die Leistungsdaten ihrer Schülerinnen und Schüler nehmen konnten.

Die Abbildung 1 stellt den Studienaufbau im Überblick dar.

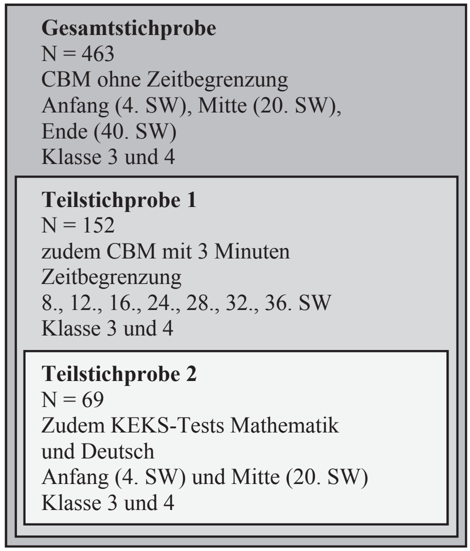


Abbildung 1: Zusammenfassung des Studienaufbaus

Anmerkungen: SW – Schulwoche

Instrumente

Zur Validierung der oben beschriebenen CBM wurden im Rahmen der vorliegenden Untersuchung KEKS-Tests (May & Bennöhr, 2013) für die Fächer Mathematik sowie Deutsch genutzt.

Die KEKS-Tests erfassen die Kernkompetenzen des jeweiligen Faches. In Mathematik werden der Mengenbegriff, das Zahlenverständnis, das Verständnis von Relationen, einfaches und komplexes Rechnen, der Umgang mit dem Stellenwertsystem sowie textbasiertes Modellieren thematisiert. Der Deutschtest umfasst die Bereiche Hörverstehen, Wortschatz, Grammatik, Leseverständnis sowie Rechtschreibung.

Für beide Fächer wurden die KEKS-Tests auf Gesamttestebene ausgewertet. Deren Reliabilitäten sind mit $\alpha > .90$ (May, Bennöhr & Berger, 2014) als hoch einzuschätzen. Die Validität der Verfahren ist durch eine hohe Deckungsgleichheit der mit KEKS gemessenen Kompetenzen und den Lehrplänen der Bundesländer gegeben. Zudem wurden jeweils ca. 400 Lehrkräfte zur Validierung der Verfahren befragt, sowohl für KEKS-Mathematik als auch für KEKS-Deutsch. Dabei ergaben sich hohe Übereinstimmungen zwischen dem mit KEKS gemessenen Kompetenzniveau und den Kompetenzeinschätzungen der Lehrkräfte (May et al., 2014).

Vorgehen

Da die vorliegende Datenmatrix eine binäre Codierung in „richtig gelöst“ bzw. „falsch gelöst“ aufweist, wurde ein dichotomes Rasch-Modell zugrunde gelegt. Die Rasch-Analysen wurden mit dem Statistikprogramm R (R Core Team, 2013) mithilfe des Pakets pairwise (Heine, 2014) durchgeführt. Die Modellpassung der Items wurde anhand ihrer geschätzten Infit-Werte beurteilt. Da die Outfit-Statistiken deutlich durch Ausreißerwerte beeinflusst werden, Infit-Werte hingegen sensitiver im Bereich mittlerer Fähigkeitsausprägungen ausfallen (Linacre, 2002), wurden in der vorliegenden Untersuchung in erster Linie die Infit-Statistiken auf Abweichungen vom Erwartungswert 1 untersucht (0.70 $\leq$ Infit $\leq$ 1.30).

Zur weiteren Analyse der Güte der Items wurden gängige Itemstatistiken (Schwierigkeit, Trennschärfe) berechnet, wobei sich die Itemschwierigkeiten σ aus den im Raschmodell nach Thurston geschätzten Schwellenwerten (aufgrund dichotomer Antwortmöglichkeiten in „richtig gelöst" und „falsch gelöst") und die Trennschärfen als punktbiseriale Korrelationen der Itemrohwerte mit dem jeweiligen CBM-Gesamtwert ergeben.

Zur Analyse der Reliabilität der Verfahren wird zum einen Cronbachs α berichtet (Gesamtstichprobe), zum anderen werden die Korrelationen zwischen zwei aufeinanderfolgenden der monatlichen Testzeitpunkte als ein Äquivalent der Retest-Reliabilität aufgefasst (Teilstichprobe 1).

Die Einschätzung der Konstruktvalidität soll anhand von Korrelationen der CBM-Daten sowie zeitgleich eingesetzter Außenkriterien (konstruktnah: KEKS Mathematik; konstruktfern: KEKS Deutsch; May & Bennöhr, 2013) ermöglicht werden. Aussagen hinsichtlich der prognostischen Validität der CBM können auf Basis von Korrelationen mit den zeitlich versetzten Außenkriterien getroffen werden. Hierbei werden die Mittelwerte der ersten drei CBM-Werte in Klasse 3 (8., 12. und 16. Schulwoche) mit den Mathematik-Werten im KEKS-Test in Beziehung gesetzt.

Die Analyse der Änderungssensibilität erfolgt auf Grundlage hierarchisch-linearer Modelle (HLM; Bryk & Raudenbush, 1992) der CBM-Daten der Teilgruppe 1 (Level 2, Schülerebene) über die Zeit (Level 1, Zeitebene). Es werden getrennte Modelle für jedes CBM und Schuljahr berechnet. Der Zeitfaktor ist auf den ersten Messzeitpunkt, jeweils in der achten Schulwoche, zentriert, wodurch der Intercept als Leistungsniveau zu Beginn eines Schuljahres interpretiert werden kann. Die Slopes weisen den Leistungsanstieg als CBM-Rohwert über eine Schulwoche aus.

Ergebnisse

Raschkonformität, Schwierigkeit und Trennschärfe der Items

Auf Grundlage der Ergebnisse der Gesamtstichprobe ohne Eingrenzung der Bearbeitungszeit zu den sechs Messzeitpunkten über beide Schuljahre sind Aussagen auf Itemebene möglich. In Tabelle 4 sind die ermittelten Kennwerte der Items für jedes CBM zusammengefasst dargestellt.

Die Skalierung der Items nach dem eindimensionalen Rasch-Modell erfolgte dabei für jede der Grundrechenoperationen über die Klassenstufen 3 und 4 hinweg. Insgesamt genügen alle Items den Anforderungen an die Infit-Statistiken (0.7 $\leq$ Infit $\leq$ 1.3; Bond & Fox, 2015) und weisen damit eine hinreichende Passung zum Rasch-Modell auf.

Der Wertebereich der Itemschwierigkeiten σ fällt erwartungskonform für jedes der erarbeiteten CBM sehr breit aus. Trägt man die Parameter geordnet nach der Position auf der Testvorlage ab, zeichnet sich ein – teilweise durch einzelne Items durchbrochener – linearer Anstieg in den zuvor nach Zahlenraum definierten Niveaustufen (vgl. Tabelle 2) ab. Allerdings muss festgehalten werden, dass Sprünge in den Schwierigkeiten von Niveau zu Niveau zu verzeichnen sind. Zwar steigen die Schwierigkeiten der Items innerhalb der theoretisch postulierten Niveaustufen (zumeist) sukzessive an, die letzten Items einer Niveaustufe fallen allerdings in der Regel etwas schwieriger aus als die ersten Items der folgenden Niveaustufe. Zur besseren Verdeutlichung des Sachverhalts werden in Abbildung 2 beispielhaft die Diagramme für die CBM der Klassenstufe 4 dargestellt.

Würden alle Itemschwierigkeiten eines CBM der Größe nach geordnet werden, ergäben die Werte – jeweils für jedes CBM – annähernd eine Gerade, d. h. die Items verteilen sich jeweils gleichmäßig auf unterschiedliche Schwierigkeitsbereiche.

Die ermittelten Trennschärfen der Items liegen in einem Intervall zwischen r_{pbis} = .12 und r_{pbis} = .71 (vgl. Tabelle 4), wobei die Kennwerte lediglich bei zwei Items des CBM Addition kleiner als r_{pbis} = .20 ausfallen. Die Lösungsraten liegen für diese beiden Items messzeitpunktübergreifend bei über 98 %.

Aufgrund des Einsatzes des Rasch-Modells ist es zudem möglich, Item- und Personenparameter auf derselben Skala abzubilden. Dadurch kann die Angemessenheit des Schwierigkeitsgrades der entwickelten Aufgaben eingeschätzt werden. Die Person-Item-Maps (vgl. Abbildung 3) zeigen die Personenparameter als Histogramme sowie die gegenübergestellten Itemparameter. Es wird deutlich, dass sich der Messbereich der Items weitgehend mit den Personenparametern deckt. Lediglich für sehr leistungsstarke Kinder fehlen dem Leistungsstand entsprechende Items (bei einer Durchführung als Powertest, d. h. ohne Beschränkung der Bearbeitungszeit). Dies gilt beim CBM Division analog für Items im unteren Leistungsbereich.

Reliabilität

Weil jeweils zu Beginn, zur Mitte und zum Ende eines jeden Schuljahres die Bearbeitungszeit der CBM in der Gesamtstichprobe nicht begrenzt wurde, sind Berechnungen hinsichtlich der internen Konsistenz der vier entwickelten CBM möglich. Die Werte liegen hierbei zwischen α = .84 und α = .92, was für die Homogenität der Verfahren spricht (vgl. Tabelle 5).

Für die Teilstichprobe 1 von 152 Kindern wurden zudem Korrelationen der jeweils aufeinanderfolgenden monatlichen Testzeitpunkte berechnet. Diese fallen mit r > .70 durchweg hoch aus (vgl. Tabelle 5).

Tabelle 4: Itemstatistiken der vier CBM

	Wertebereich der Itemschwierigkeiten σ (min - max)	**mittlere Trennschärfe r_{pbis} Ø (min - max)**	**mittlerer Infit MnSq Ø (min - max)**
CBM AD	-3.67 - 4.50	0.46 (0.12 - 0.72)	0.85 (0.70 - 1.02)
CBM SU	-3.33 - 3.23	0.56 (0.22 - 0.71)	0.88 (0.75 - 1.05)
CBM MU	-5.19 - 4.54	0.56 (0.33 - 0.67)	0.89 (0.71 - 1.20)
CBM DI	-4.69 - 3.54	0.59 (0.44 - 0.70)	0.90 (0.70 - 1.23)

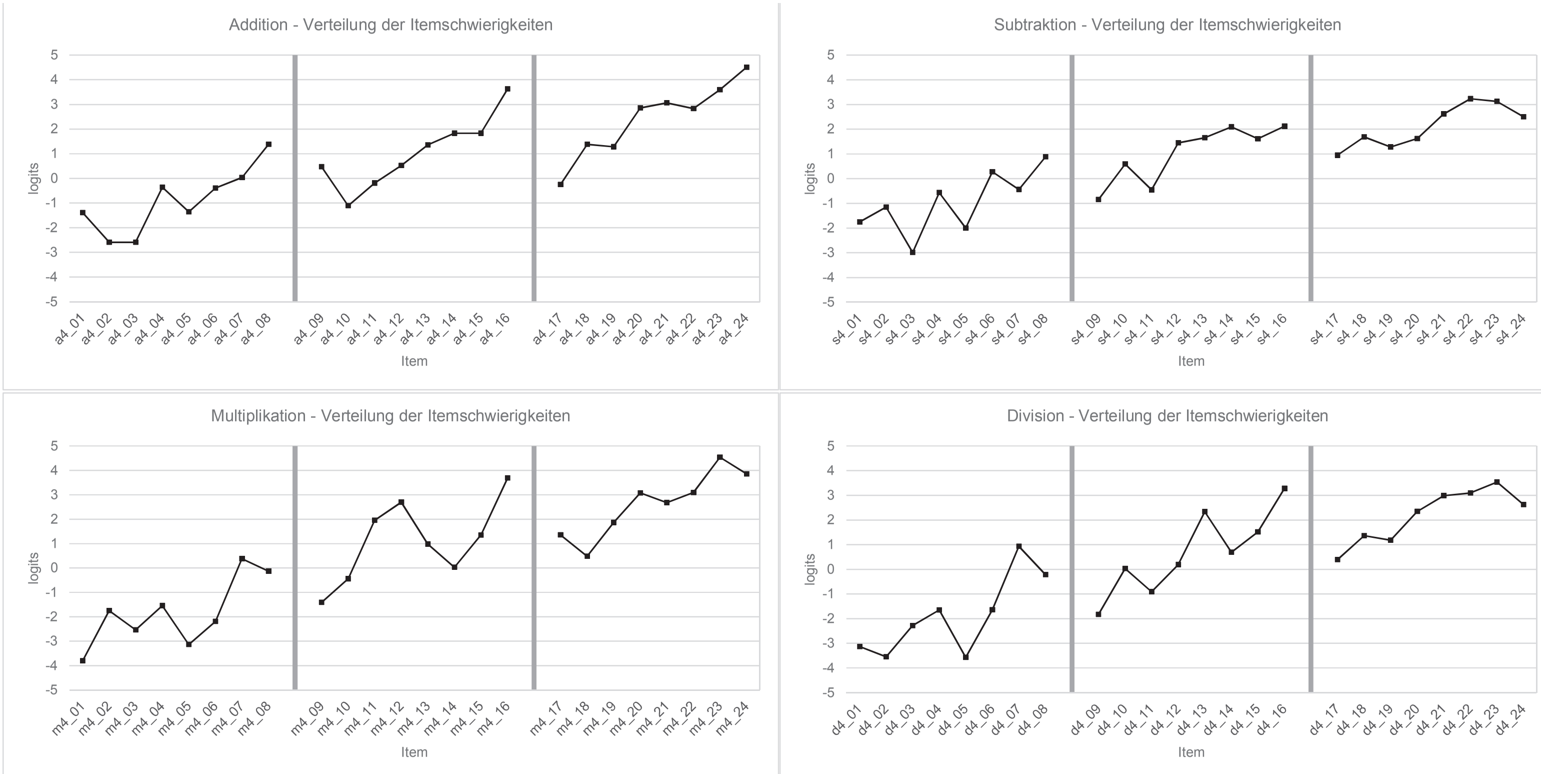
Anmerkungen: AD – Addition, SU – Subtraktion, MU – Multiplikation, DI – Division

Tabelle 5: Cronbachs a in der Gesamtstichprobe und Korrelationen der CBM-Daten zweier aufeinanderfolgender monatlicher Messzeitpunkte in der Teilstichprobe 1

	mittleres Cronbachs α Ø (min - max)	**mittlere Korrelation aufeinanderfolgender Messzeitpunkte in Klasse 3 Ø (min - max)**	**mittlere Korrelation aufeinanderfolgender Messzeitpunkte in Klasse 4 Ø (min - max)**
CBM AD	α = .86 (.84 - .88)	r = .77 (.74 - .80)	r = .76 (.70 - .83)
CBM SU	α = .91 (.89 - .92)	r = .79 (.72 - .85)	r = .83 (.80 - .86)
CBM MU	α = .89 (.85 - .91)	r = .78 (.73 - .82)	r = .79 (.77 - .82)
CBM DI	α = .90 (.89 - .91)	r = .78 (.73 - .83)	r = .81 (.79 - .83)

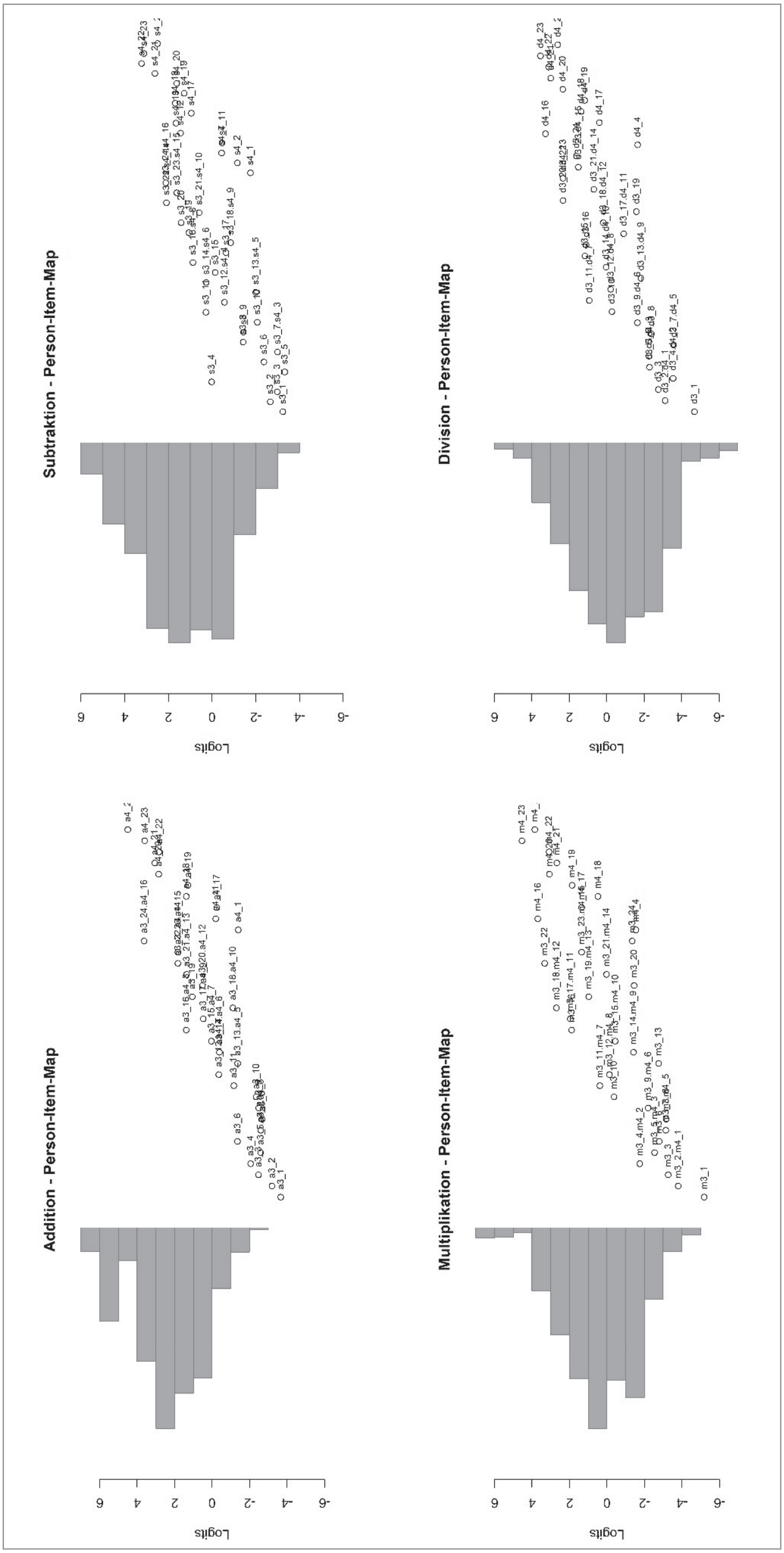
Anmerkungen: AD – Addition, SU – Subtraktion, MU – Multiplikation, DI – Division

Abbildung 2: Verteilung der Itemschwierigkeiten am Beispiel der entwickelten CBM für Klasse 4



Anmerkungen: Die grauen vertikalen Markierungen weisen die nach Zahlenraum definierten Niveaustufen in den CBM aus (vgl. Tabelle 2)

Abbildung 3: Person-Item-Maps der entwickelten CBM



Validität

Die Interkorrelationen der vier CBM Addition, Subtraktion, Multiplikation und Division über die drei Messzeitpunkte je Schuljahr (Gesamtstichprobe) variieren in einem Bereich zwischen r = .55 und r = .85 bei einem Mittelwert von r = .75. Erwartungskonform sind demnach hohe Zusammenhänge zwischen den Ergebnissen der einzelnen CBM festzuhalten.

Die Teilstichprobe 2 (N = 69) hat jeweils zur Mitte und zum Ende der Schuljahre den Mathematik- und Deutschtest des KEKS-Diagnosekonzepts bearbeitet. Werden diese Ergebnisse mit zeitgleich eingesetzten CBM-Testzeitpunkten korreliert, können Einschätzungen der Konstruktvalidität vorgenommen werden. Erwartungskonform fallen die Korrelationen mit den konstruktnahen Verfahren (Mathematiktests) insgesamt hoch aus (r ≥ .50) und höher als die Zusammenhangsmaße mit den konstruktfernen Deutschtests. Entgegen den aufgestellten Erwartungen ergeben sich jedoch zum Teil auch zwischen den CBM-Ergebnissen und den Deutsch-Werten in KEKS 3 hohe Korrelationen (bis r = .53).

Zur Einschätzung der prognostischen Validität der CBM wurden die Mittelwerte der ersten drei Messungen der Teilstichprobe 2 in Klasse 3 gebildet und mit den zeitlich versetzt erhobenen Mathematikleistungen im KEKS-Test zu Beginn und zur Mitte der vierten Klasse korreliert (vgl. Tabelle 7). Insgesamt sind eher hohe, zumindest aber mittlere Zusammenhänge zu verzeichnen. Auffällig ist, dass die Zusammenhangsmaße zwischen den CBM und den zur Mitte des vierten Schuljahres erhobenen Mathematikleistungen höher ausfallen (r zwischen .52 und .67) als mit denen zu Beginn des vierten Schuljahres (r zwischen .38 und .56).

Änderungssensibilität

Zur Analyse der Veränderung der CBM-Daten über die Zeit wurden auf Basis der Da-

Tabelle 6: Korrelationen mit den Außenkriterien

	Außenkriterium	**CBM AD**	**CBM SU**	**CBM MU**	**CBM DI**
Anfang Klasse 3	KEKS 3 Mathe	.63**N = 62	.66**N = 62	.55**N = 62	.70**N = 62
	KEKS 3 Deutsch	.53**N = 63	.34**N = 63	.42**N = 63	.49**N = 63
Mitte Klasse 3	KEKS 3 Mathe	.60**N = 67	.67**N = 67	.61**N = 69	.68**N = 69
	KEKS 3 Deutsch	.52**N = 63	.45**N = 63	.50**N =65	.40**N = 65
Anfang Klasse 4	KEKS 4 Mathe	.64**N = 68	.59**N = 68	.58**N = 67	.59**N = 67
	KEKS 4 Deutsch	.31*N = 66	.40**N = 66	.47**N = 65	.39**N = 65
Mitte Klasse 4	KEKS 4 Mathe	.59**N = 50	.71**N = 50	.80**N = 50	.80**N = 50
	KEKS 4 Deutsch	.44**N = 50	.30*N = 50	.34*N = 50	.30*N = 50

Anmerkungen: AD – Addition, SU – Subtraktion, MU – Multiplikation, DI – Division, * – p < .05; ** – p < .01

Tabelle 7: Korrelationen der Mittelwerte der ersten drei CBM-Werte in Klasse 3 mit später eingesetzten Außenkriterien

Außenkriterium	**CBM AD**	**CBM SU**	**CBM MU**	**CBM DI**
KEKS 4 Mathematik Anfang	.38**N = 68	.52**N = 68	.44**N = 68	.56**N = 68
KEKS 4 Mathematik Mitte	.52**N = 63	.63**N = 63	.52**N = 63	.67**N = 63

Anmerkungen: AD – Addition, SU – Subtraktion, MU – Multiplikation, DI – Division, ** – p < .01

ten der Teilstichprobe 1 Mehrebenenmodelle (HLM) für alle CBM jeweils für die Klasse 3 und die Klasse 4 spezifiziert (vgl. Tabelle 8). Die Ergebnisse der HLM weisen zu Beginn der Schuljahre durchschnittliche CBM-Werte von β_{00} = 6.95 bis β_{00} = 15.17 (Klasse 3; jeweils $p < .001$) bzw. β_{00} = 9.31 bis β_{00} = 15.19 (Klasse 4; jeweils $p < .001$) aus. Die höchsten Summenscores werden jeweils hinsichtlich der Addition erzielt, es folgen die der Subtraktion, Multiplikation und Division.

Zur Einschätzung der Änderungssensibilität sind insbesondere die mittleren Anstiege der CBM-Daten über die Zeit von Bedeutung. Aus den HLM resultieren durchschnittliche Steigungskoeffizienten von β_{10} = 0.10 bis β_{10} = 0.14 je Schulwoche für Klasse 3 (jeweils $p < .001$) bzw. β_{10} = 0.07 bis β_{10} = 0.12 je Schulwoche für Klasse 4

Tabelle 8: Ergebnisse der random-coefficient-Modelle zur Analyse der CBM-Daten über die Zeit (Teilstichprobe 1)

		CBM Klasse 3			**CBM Klasse 4**		
Feste Effekte		**Koeff (SE)**	**t**	**df**	**Koeff (SE)**	**t**	**df**
Modelle für die intercepts							
Niveau, β_{00}	AD	15.17 (0.36)***	42.49	151	15.19 (0.35)***	42.45	151
	SU	11.52 (0.40)***	28.73	151	11.69 (0.43)***	27.10	151
	MU	9.81 (0.33)***	29.68	151	10.78 (0.32)***	34.10	151
	DI	6.95 (0.34)***	20.72	151	9.31 (0.37)***	25.55	151
Modelle für die slopes							
Anstieg, β_{10}	AD	0.11 (0.01)***	11.32	151	0.07 (0.01)***	7.25	151
	SU	0.13 (0.01)***	12.50	151	0.10 (0.01)***	9.85	151
	MU	0.14 (0.01)***	14.93	151	0.12 (0.01)***	14.61	151
	DI	0.10 (0.01)***	10.52	151	0.12 (0.01)***	14.10	151
Zufallseffekte		**Var**	**$\chi 2$**	**df**	**Var**	**$\chi 2$**	**df**
Niveau u_{0i}	AD	17.36***	1380.95	151	16.61***	1359.48	151
	SU	22.15***	1499.10	151	25.89***	1782.95	151
	MU	14.35***	1068.05	151	13.36***	1198.27	151
	DI	14.61***	1002.53	151	18.13***	1422.18	151
Anstieg u_{1i}	AD	0.006***	285.02	151	0.008***	302.03	151
	SU	0.008***	296.90	151	0.077***	260.63	151
	MU	0.005***	230.92	151	0.003**	208.80	151
	DI	0.004***	231.11	151	0.004***	232.25	151
Level-1-Fehler e_{ij}	AD	4.72			4.45		
	SU	5.44			5.14		
	MU	5.12			4.10		
	DI	5.69			4.61		
Devianz (df)	AD	5129.56 (4)			4660.65 (4)		
	SU	5128.97 (4)			4833.40 (4)		
	MU	4988.48 (4)			4533.87 (4)		
	DI	5177.65 (4)			4684.89 (4)		

Anmerkungen: ** - $p < .01$; *** - $p < .001$

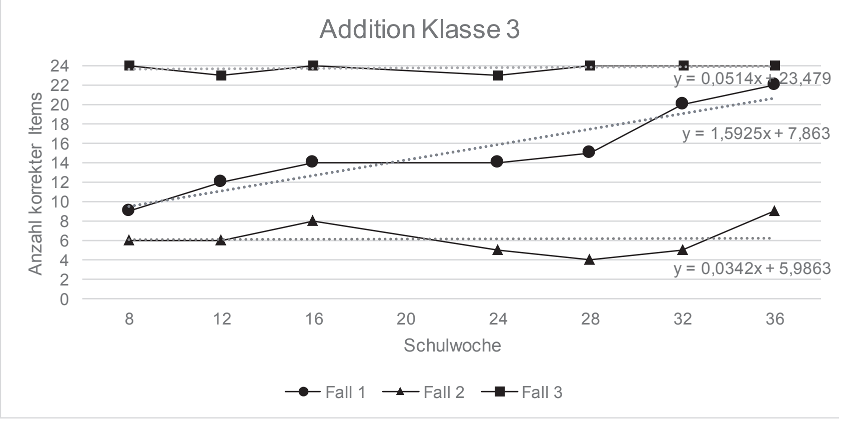
(jeweils $p < .001$). Das bedeutet, dass im Mittel – je nach CBM – nach 7 bis 14 Schulwochen von einem durchschnittlichen Anstieg von einem Rohwertpunkt ausgegangen werden kann.

Die Zufallseffekte weisen für alle CBM weitere schülerspezifische Unterschiede im Niveau und Anstieg der CBM-Daten aus, die im Rahmen der gerechneten Modelle nicht erklärt werden konnten.

Nach Tymms (2004) können durch Verrechnung der in den HLM berichteten slopes bzw. deren Standardfehler mit dem Level-1-Fehler Effektstärkenintervalle für diese Koeffizienten ermittelt werden, die wie Cohens d interpretierbar sind. Hochgerechnet auf ein Schuljahr mit 40 Schulwochen ergeben sich in den Daten folgende Effektstärkenintervalle für die einzelnen CBM und Schuljahre:

- Addition Klasse 3: $d = 0.93 \pm 0.08$
- Subtraktion Klasse 3: $d = 0.95 \pm 0.07$
- Multiplikation Klasse 3: $d = 1.09 \pm 0.07$
- Division Klasse 3: $d = 0.70 \pm 0.07$
- Addition Klasse 4: $d = 0.62 \pm 0.08$
- Subtraktion Klasse 4: $d = 0.77 \pm 0.07$
- Multiplikation Klasse 4: $d = 1.17 \pm 0.09$
- Division Klasse 4: $d = 1.04 \pm 0.08$

Zwar werden individuelle Differenzen in der Entwicklung der Kinder im Rahmen der hierarchisch-linearen Modellierung berücksichtigt, die oben berichteten Werte beziehen sich jedoch jeweils auf die Gruppe der



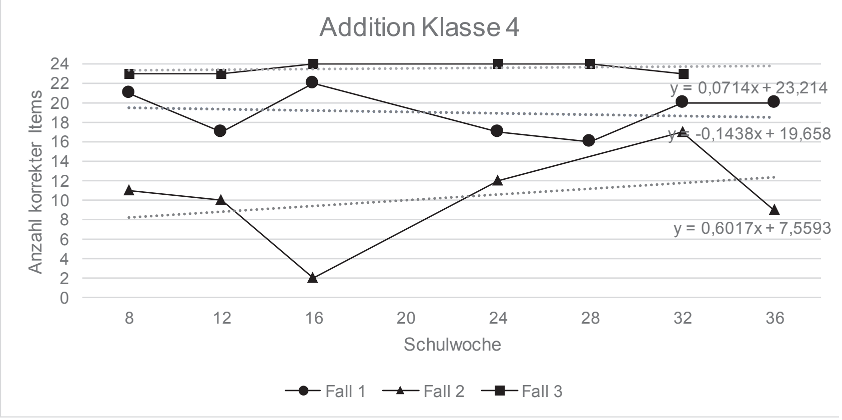


Abbildung 4: Datenverlauf im Einzelfall am Beispiel Addition in Klasse 3 und 4

Kinder der Teilstichprobe 1 als Gesamtes. Um differenziertere Aussagen zu den individuellen Lernverläufen treffen zu können, wurden exemplarisch für den Bereich Addition drei Kinder im Einzelfall analysiert (vgl. Abbildung 4). Hier zeigt sich die bereits bei Strathmann und Klauer (2010) bzw. Klauer (2011) berichtete Spannbreite an verschiedenen Leistungsverläufen. So weist Fall 1 einen relativ hohen vierwöchentlichen Anstieg der CBM-Werte im dritten Schuljahr auf, welcher im vierten Schuljahr zu einem leicht negativen Wert umkehrt. Im Fall 2 liegt der Anstieg der Leistung im Laufe der dritten Klasse nahe Null, in der vierten Klasse ist hier jedoch ein höherer Leistungsanstieg zu verzeichnen. Fall 3 weist einen Deckeneffekt über die gesamte Untersuchungszeit auf. Dieses Kind erzielt bereits jeweils zum Start der Schuljahre Ergebnisse nahe oder gleich der Maximalpunktzahl der CBM.

Diskussion

Das Ziel des vorliegenden Beitrags lag in der Darstellung der Entwicklung von CBM, die sich auf die Rechenfertigkeiten von Dritt- und Viertklässlern in den Grundrechenoperationen beziehen, sowie in der Analyse von deren psychometrischer Güte. Hierzu wurden über zwei Schuljahre hinweg jeweils dreimal die CBM-Daten einer Kindergruppe erhoben (N = 463). Von 152 dieser Schülerinnen und Schüler wurden zudem monatliche CBM-Daten erhoben. Für 69 Kinder dieser Gruppe liegen zudem Werte von konvergenten und diskriminanten Außenkriterien vor.

Um die Inhaltsvalidität zu gewährleisten, wurde auf Basis fachdidaktischer Stufungen der Grundrechenoperationen eine Stratifizierung der Items nach schwierigkeitsgenerierenden Kriterien vorgenommen (unterschiedliche Zahlenräume, Anzahl zu verrechnender Stellen, Einsatz unterschiedlicher Aufgabenformate). Es wurden hierbei Aussagenformen definiert, die jeweils diskrete Aufgabenmengen umschließen (Klauer, 2011). Dabei wurde darauf geachtet, die Eingrenzungsbedingungen immer so eng zu fassen, dass die resultierenden Itempools als homogene Aufgabenklassen angesehen werden können.

Bis auf wenige Ausnahmen (zwei Items weisen eine zu geringe Trennschärfe bei sehr hoher Lösungswahrscheinlichkeit auf) erscheinen die Items geeignet, den Lernverlauf der Dritt- bzw. Viertklässler zu erfassen. So genügen alle Aufgaben den angelegten Kriterien eines eindimensionalen Rasch-Modells. Die Schwierigkeiten der Items weisen in allen CBM hohe Spannbreiten auf, was wünschenswert ist, da sich die Verfahren somit für den Einsatz bei Kindern mit unterschiedlichen Fähigkeitsausprägungen eignen. Dies wird auch in den Person-Item-Maps deutlich, so deckt sich der Messbereich der Items weitgehend mit dem Leistungsbereich der Stichprobe. Lediglich im sehr hohen Leistungsbereich (bei Durchführung als Powertest) und bei der Division im unteren Leistungsbereich fehlt es an geeigneten Items. Diesbezüglich ist anzumerken, dass ein wesentliches Ziel im Einsatz von CBM die frühzeitige Identifikation von problematischen Entwicklungen darstellt, woraus insbesondere eine hinreichende Differenzierungsfähigkeit im unteren Leistungsbereich resultiert. Insofern ist für das CBM Division eine Erweiterung der Items in diesem Bereich angezeigt.

Die internen Konsistenzen wurden bei unbegrenzter Bearbeitungszeit zu sechs Messzeitpunkten ermittelt und liegen durchweg im hohen Bereich (α = .84 - .92). Methodenkritisch muss jedoch die Übertragbarkeit der so berechneten Homogenität infrage gestellt werden, da die CBM in der Praxis mit einer Zeitbegrenzung bearbeitet werden. Werden die CBM als Speed-Test durchgeführt, liegen die berechneten Korrelationen aufeinanderfolgender Testzeitpunkte bei r = .70 - .86. Streng genommen handelt es sich hierbei jedoch nicht um eine Retest-Reliabilität, da nicht ein- und dasselbe Verfahren erneut durchgeführt wurde.

Ebenfalls kann nicht von einer „echten" Paralleltestreliabilität gesprochen werden, da die Verfahren jeweils um vier Schulwochen versetzt bearbeitet wurden. Aufgrund der engen Definition der Itempositionen und der daraus resultierenden gleichen Struktur der CBM-Versionen, kann jedoch – insgesamt betrachtet – von einer ausreichenden Parallelität der Verfahren ausgegangen werden (vgl. auch Klauer, 2011). Zusammenfassend ist zudem darauf hinzuweisen, dass die ermittelten Retest-Reliabilitäten geringer ausfallen als die Kennwerte der internen Konsistenz, was nach Klauer (2011) Kennzeichen eines änderungssensiblen Verfahrens ist.

Setzt man die empirisch gewonnenen Itemschwierigkeiten in Bezug zur vorab theoretisch vorgenommenen Reihung, ist zu erkennen, dass für die CBM als jeweils Ganzes nicht von einer gleichmäßigen Steigung der Itemschwierigkeiten ausgegangen werden kann, wohl aber weitgehend innerhalb der nach Zahlenraum definierten Niveaustufen innerhalb der CBM (vgl. Abbildung 2). Die Befunde machen deutlich, dass eine ausschließliche Orientierung an Zahlenräumen für die Definition von Schwierigkeitsklassen nicht hinreichend ist. So erscheint es logisch, dass bspw. die Aufgabe 200 + 100 = _ (Niveau Klasse 3) nicht schwieriger ist als 57 + 38 = _ (Niveau Klasse 2). Dies spricht für das gewählte Vorgehen, zur Generierung verschiedener Schwierigkeitsgrade zusätzlich die zur Ergebnisfindung zu verarbeitenden Stellen sowie verschiedene Aufgabenformate zu berücksichtigen. Aufgrund der punktuellen Diskrepanzen zwischen theoretisch definierter und empirisch ermittelter Schwierigkeit der Aufgaben, sollte über eine angepasste Reihung innerhalb der Niveaustufen nachgedacht werden.

Hierbei handelt es sich aber um eine grundsätzliche Frage. Für Speed-Tests wird eine aufsteigende Schwierigkeit der Items gefordert (Lienert & Raatz, 1998), die Aufgaben müssten somit strikt nach empirischen Kriterien angeordnet werden. Resultieren würde ein Verfahren, in dem Aufgaben aus den verschiedenen Klassenstufen im Wechsel bearbeitet würden. Für das hier gewählte Vorgehen, inhaltsvalide Aufgabengruppen für jede Klassenstufe zu generieren und diese nacheinander vorzulegen, sprechen inhaltliche Überlegungen. Die Aufgaben eines Niveaus repräsentieren die Teillehrziele der jeweiligen Klassenstufe. So werden beispielsweise in dem CBM zur Addition der dritten Klasse zunächst Aufgaben mit ganzen Hundertern, anschließend Aufgaben mit Hundertern und Zehnern bzw. Hundertern und Einern und schließlich Aufgaben mit Hundertern, Zehnern und Einern bearbeitet. Diese sukzessive Steigerung der Anforderungen spiegelt die Reihenfolge der unterrichtlichen Behandlung der Addition in der dritten Klassenstufe wider. Durch ebendiese Anordnung der Aufgaben beinhaltet das Testergebnis auch eine qualitative Aussage über die Schülerleistung in Bezug auf die curricularen Anforderungen der jeweiligen Klassenstufe. So könnten im Sinne einer Kriteriumsorientierung beispielsweise zwei richtig gelöste Aufgaben als rudimentäre, vier als grundlegend und sechs als hinreichend verfügbare Kompetenzen innerhalb der Niveaustufe interpretiert werden. Diese Möglichkeit der Verortung der Kinder im Kompetenzkontinuum steigert die Förderrelevanz des Verfahrens maßgeblich.

Aus demselben Grund wurde bei den vorliegenden CBM beschlossen, Verfahren für jede Grundrechenoperation zu entwickeln. In diesem Punkt unterscheidet sich das hier gewählte Vorgehen von bisherigen CBM für das höhere Grundschulalter (Strathmann & Klauer, 2012). Zwar sprechen die in den Interkorrelationsanalysen ermittelten hohen Zusammenhänge für eine große strukturelle Ähnlichkeit der Anforderungen in den vier Grundrechenarten, allerdings ist nicht unbedingt davon auszugehen, dass ein Kind, welches beispielsweise gute Fertigkeiten im Addieren zeigt, auch notwendigerweise gut multiplizieren oder dividieren kann. Schließlich handelt es sich um verschiedene Operationen, in denen die Entwicklungsstände durchaus unterschiedlich sein können. So könnte es sein, dass das beschriebene Kind in der Addition einen altersgerechten Leistungsstand aufweist, den curricularen Anforderungen in der Multiplikation oder Division seiner Klassenstufe hingegen nicht entspricht. Durch die vorliegenden CBM würde, bei einem Zeitaufwand von viermal drei Minuten Bearbeitungszeit, ein solches Leistungsprofil sichtbar werden, woraus entsprechende Konsequenzen für die Förderung abgeleitet werden können.

Um weitere Hinweise zur Konstruktvalidität der entwickelten CBM zu erhalten, wurden neben Interkorrelationsanalysen auch solche mit konvergenten sowie diskriminanten Verfahren durchgeführt. Zu diesem Zweck hat die Teilstichprobe 2 die Mathematik- bzw. Deutschtests des KEKS-Diagnosekonzeptes bearbeitet. Die hohen Korrelationen mit den Mathematiktests sowie die geringeren Zusammenhänge mit den Tests der Deutsch-Reihe sind Kennzeichen für die konvergente und diskriminante Validität der erarbeiteten CBM. Dass es durchaus hohe Übereinstimmungen von Mathematikleistungen und schriftsprachlichen Aspekten gibt, deckt sich mit entsprechenden Befunden von Krajewski, Schneider und Nieding (2008) sowie mit denen von Validierungsstudien zu Testverfahren mit ähnlichem Anwendungsbereich (Sikora, 2017).

Die Zusammenhänge der Ergebnisse in den entwickelten CBM mit zeitlich später eingesetzten Verfahren fallen durchgängig signifikant aus und liegen im mittleren bis zumeist hohen Bereich. Diese können als Indikatoren für eine gute Voraussagekraft der entwickelten CBM angesehen werden. Schwer erklärbar sind hingegen die durchgängig höheren Korrelationen mit dem KEKS-Mathematiktest zur Schuljahresmitte der vierten Klassenstufe als mit denen zum Schuljahresbeginn, welche entgegen den Erwartungen ausfallen. Da es keine Unterschiede im zur Validierung genutzten Testverfahren gibt, muss vermutlich von Verzerrungen aufgrund des geringen Stichprobenumfangs ausgegangen werden. Damit wäre eine grundsätzliche Schwäche der vorliegenden Untersuchung angesprochen: Insbesondere bei den Analysen zur Validität fallen die Stichprobenumfänge sehr gering aus, was die Aussagekraft der Daten erheblich mindert. Die diesbezüglichen Ergebnisse sollten als erste Hinweise zur Validität der entwickelten CBM gedeutet werden. Entsprechende Folgeuntersuchungen sind angezeigt.

Zwar wirken die im Rahmen der HLM ermittelten slopes mit Werten zwischen $\beta_{10} = 0.07$ und $\beta_{10} = 0.14$ zunächst eher gering, hierbei muss jedoch bedacht werden, dass diese Werte den Lernfortschritt der untersuchten Gruppe innerhalb einer Schulwoche auf Rohwertebene anzeigen. Rechnet man die Koeffizienten hoch, wird deutlich, dass – auf die Gesamtgruppe bezogen – bereits nach 7 bis 14 Schulwochen Änderungen im Lernverlauf erkennbar werden. Damit reagieren die vorliegenden CBM, je nach Grundrechenart und Klassenstufe, ähnlich sensibel auf Kompetenzänderungen wie bereits etablierte Verfahren (Strathmann & Klauer, 2012). Zur Steigerung der Änderungssensibilität könnte darüber nachgedacht werden, die entwickelten CBM ohne Zeitbegrenzung durchzuführen. Dabei würde vermutlich eine höhere Variabilität der Ergebnisse resultieren, weil die Kinder mehr Möglichkeiten zur Entwicklung hätten.

Auf Individualebene zeichnen sich bei den vorgestellten CBM, auch bei einer Durchführung mit Zeitbegrenzung, sehr wahrscheinlich bereits nach wenigen Testzeitpunkten sichtbare Lernfortschritte ab. Bei ausbleibenden Leistungszuwächsen über einen Zeitraum von etwa 10 Schulwochen sollte somit über eine Modifikation des Unterrichts nachgedacht werden, sofern die schüler- und situationsspezifischen Informationen zu den CBM-Ergebnissen korrespondieren.

Werden die hier berichteten wöchentlichen Anstiege auf ein Schuljahr (im Durchschnitt 40 Schulwochen) extrapoliert, resultieren Effektstärken, die mit empirisch gewonnenen Erkenntnissen über zu erwartende Lernzuwächse vergleichbar sind. So berichten Reiss und Winkelmann (2009) eine durchschnittliche Effektstärke von d = 0.84 für die Kompetenzentwicklung im Fach Mathematik von der dritten zur vierten Jahrgangsstufe. In diesem Zusammenhang ist jedoch darauf hinzuweisen, dass die Lehrkräfte innerhalb der vorliegenden Studie über eine Internetplattform (Voß et al., 2016) fortlaufend den aktuellen Leistungsstand auf Grundlage der CBM-Daten einsehen konnten. Dies kann die pädagogische Arbeit dahingehend beeinflusst haben, dass unterrichtliche Überlegungen entsprechend den sich abzeichnenden Leistungsprofilen der Kinder angepasst wurden. In diesem Fall würden die hier dargestellten Effekte tendenziell überschätzt.

Ähnlich wie bei Klauer und Strathmann (2010) bzw. Klauer (2011) zeigt sich auch in den vorliegenden Daten, dass nicht alle Kinder tatsächlich Lernfortschritte über die Zeit erzielen. Exemplarisch für den Bereich Addition wurden die CBM-Ergebnisse dreier Kinder über das dritte und vierte Schuljahr dargestellt (vgl. Abbildung 4). Dabei wurden Lernplateaus ebenso deutlich wie starke Leistungssteigerungen. Da die Lehrkräfte der teilnehmenden Klassen die Leistungsentwicklungen der Kinder ihrer Klasse fortlaufend über eine Internetplattform (Voß et al., 2016) einsehen konnten, stellt sich die Frage, warum es zum Teil dennoch zu stagnierenden Verläufen kommt. Wie schwer es Lehrkräften fällt, Kinder entsprechend ihren Lernvoraussetzungen mathematisch zu fördern, wird in der mathematikdidaktischen Fachliteratur immer wieder beschrieben. Relativ einstimmig wird den Praktikerinnen und Praktikern unsystematisches Vorgehen und eine zu geringe Passung zwischen Lernvoraussetzungen und Förderung vorgeworfen (u. a. Wielpütz, 2010; Zimmermann, 2011; Häsel-Weide, Nührenbörger, Moser Opitz & Wittich, 2015). Diese „didaktogene Komponente“ (Gaidoschik, 2003, S. 125) von mathematischen Lernschwierigkeiten wird verstärkt durch den Anspruch, einer Vielzahl, offenkundig sehr unterschiedlicher, Leistungsprofile der Kinder einer Klasse gerecht werden zu müssen. Zudem gestalten sich die Lernverläufe von Kindern ganz und gar nicht eindeutig. Lineare Zuwächse, wie sie oben beschrieben werden, sind eher selten. In der Regel hat man es mit gezackten Verläufen zu tun, die – mathematisch gesehen – nur mit Polynomen höherer Ordnungen hinreichend beschrieben werden können (Stathmann & Klauer, 2010). Ebendiese Leistungskurven sind schwer fassbar und entsprechend schwer zu interpretieren.

In diesem Zusammenhang stellen sich die folgenden Fragen: Können die Lehrkräfte die CBM-Daten richtig deuten? Was folgt aus den CBM-Ergebnissen? Schließt sich Förderung an? Ist diese Förderung adäquat auf die Problemlage des Kindes abgestimmt? Wie muss Feedback aussehen, um Lehrkräfte in ihrer pädagogischen Tätigkeit zu unterstützen?

Die Datenlage der vorliegenden Untersuchung lässt dahingehend keine Aussagen zu, jedoch liegt auf ebenjenen Fragen ein zukünftiger Forschungsschwerpunkt der Autoren.

Literatur

Bond, T. G. & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3. Aufl.). New York & London, UK: Routledge.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

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

Deno, S. L. (2003). Developments in Curriculum-based Measurement. *Journal of Special Education, 37*, 184–192.

Fritz, A., Ricken, G. & Gerlach, M. (2007). *Kalkulie. Handreichung zur Durchführung einer Diagnose*. Berlin: Cornelsen.

Fuchs, L. S. (2004). The Past, Present, and Future of Curriculum-Based Measurement Research. *School Psychology Review, 33*, 188-192.

Gaidoschik, M. (2003). Rechenstörungen: Die „didaktogene Komponente“. Kritische Thesen zur „herkömmlichen Unterrichtspraxis“ in drei Kernbereichen der Grundschulmathematik. In F. Lenart, N. Holzer & H. Schaupp (Hrsg.), *Rechenschwäche – Rechenstörung – Dyskalkulie: Erkennung, Prävention, Förderung* (S. 125-153). Graz, AT: Leykam.

Gebhardt, M. F., Schwab, S., Schaupp, H., Rossmann, P. & Gasteiger Klicpera, B. (2012). Heterogene Gruppen in mathematischen Grundfertigkeiten: Eine explorative Erkundung der Fähigkeiten im Grundrechnen in Integrationsklassen der 5. Schulstufe. *Zeitschrift für Inklusion*. Zugriff am 28.05.2017. Verfügbar unter http://www.inklusion-online.net/index.php/inklusion-online/article/view/69.

Grassmann, M., Eichler, K.-P., Mirwald, E. & Nitsch, B. (2014). *Mathematikunterricht. Kompetent im Unterricht der Grundschule. 3. korrigierte und veränderte Auflage*. Baltmannsweiler: Schneider Verlag Hohengehren.

Hammer, S., Reiss, K., Lehner, M. C., Heine, J.-H., Sälzer C. & Heinze, A. (2015). Mathematische Kompetenz in PISA 2015: Ergebnisse, Veränderungen und Perspektiven. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. & O. Köller (Hrsg.), *Eine Studie zwischen Kontinuität und Innovation* (S. 220-247). Münster: Waxmann.

Häsel-Weide, U., Nührenbörger, M., Moser Opitz, E. & Wittich, C. (2015). Ablösung vom zählenden Rechnen. Fördereinheiten für heterogene Lerngruppen. 3. Auflage. Seelze: Klett-Kallmeyer.

Hartmann, E. & Müller, C. (2014). *Lernfortschrittsdiagnostik: Grundrechenarten*. Hornburg: Persen.

Hattie, J. (2013). *Lernen sichtbar machen*. Baltmannsweiler: Schneider-Verl. Hohengehren.

Heine, J.-H. (2014). *pairwise: Rasch Model Parameters by Pairwise Algorithm* [Computer software]. Munich. Zugriff am 28.05.2017. Verfügbar unter http://cran.r-project.org/web/packages/pairwise/index.html (R package version 0.4.1).

Käpnick, F. (2014). *Mathematiklernen in der Grundschule*. Berlin: Springer.

Klauer, K. J. (1987). *Kriteriumsorientierte Tests*. Göttingen: Hogrefe.

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

Krajewski, K., Schneider, W. & Nieding, G. (2008). Zur Bedeutung von Arbeitsgedächtnis, Intelligenz, phonologischer Bewusstheit und früher Mengen-Zahlen-Kompetenz beim Übergang vom Kindergarten in die Grundschule. *Psychologie in Erziehung und Unterricht, 55* , 118-131.

Krauthausen, G. & Scherer, P. (2007). *Einführung in die Mathematikdidaktik. 3. Auflage*. München: Elsevier.

Kultusministerkonferenz (2005). *Bildungsstandards im Fach Mathematik für den Primarbereich*. München: Luchterhand.

Lienert, G. A. & Ratz, U. (1998). *Testaufbau und Testanalyse. 6. Auflage*. Weinheim: Beltz.

Linacre, J. M. (2002). What do Infit and Outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*, 878.

May, P. & Bennöhr, J. (2013). *Kompetenzerfassung in Kindergarten und Schule (KEKS). Handbuch*. Berlin: Cornelsen.

May, P., Bennöhr, J. & Berger, C. (2014). Lernentwicklungsmonitoring mit KEKS. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), Lernverlaufsdiagnostik. *Tests und Trends, N.F. Bd. 12* (S. 257-280). Göttingen: Hogrefe.

Mislevy, R. J., Beaton, A. E., Kaplan, B. & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133-161.

R Core Team (2013). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing* [Computer software]. Vienna, Austria. Retrieved from http://www.R-project.org.

Reiss, K. & Winkelmann, H. (2009). Kompetenzstufenmodelle für das Fach Mathematik im Primarbereich. 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. 120-141). Weinheim: Beltz.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion. 2. vollständig überarbeitete und erweiterte Auflage*. Bern, CH: Huber.

Strathmann, A. M. & Klauer, K. J. (2010). Lernverlaufsdiagnostik: Ein Ansatz zur längerfristigen Lernfortschrittsmessung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 42*, 111-122.

Strathmann, A. M. & Klauer, K. J. (2012). *Lernverlaufsdiagnostik Mathematik. LVD-M 2-4*. Göttingen: Hogrefe.

Schipper, W. (2009). *Handbuch für den Mathematikunterricht an Grundschulen*. Braunschweig: Schroedel.

Selter, C., Walter, D., Walther, G., & Wendt, H. (2016). Mathematische Kompetenzen im internationalen Vergleich: Testkonzeption und Ergebnisse. In H. Wendt, W. Bos, C. Selter, O. Köller, K. Schwippert & D. Kasper (Hrsg.), *TIMSS 2015: Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern im internationalen Vergleich* (S. 79-136). Münster: Waxmann.

Sikora, S. (2017). *Lernverlaufsdiagnostik im Mathematikunterricht. Theoretische Grundlagen, Konzeption und Güte eines formativen Schulleistungstests für dritte Klassen*. Hamburg: Dr. Kovač.

Tymms, P. (2004). Effect sizes in multilevel models. In I. Schagen & K. Elliot (Eds.), *But what does it mean? The use of effect sizes in educational research* (p. 55-66). Slough, UK: National Foundation for Educational Research.

Voß, S. & Hartke, B. (2014). Curriculumbasierte Messverfahren (CBM) als Methode der formativen Leistungsdiagnostik im RTI-Ansatz. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (Tests & Trends, NF Bd. 12, S. 83-99). Göttingen: Hogrefe.

Voß, S. (2016). Rechengeschwindigkeit, -präzision oder -flüssigkeit? Zur Vorhersage und Förderung der Rechenleistungen von Erstklässlern. *Heilpädagogische Forschung, 42*, 13-24.

Voß, S., Blumenthal, Y., Mahlau, K., Marten, K., Diehl, K., Sikora, S. & Hartke, B. (2016). *Der Response-to-Intervention-Ansatz in der Praxis. Evaluationsergebnisse zum Rügener Inklusionsmodell*. Münster: Waxmann.

Voß, S., Sikora, S. & Hartke, B. (2017). Lernverlaufsdiagnostik als zentrales Element der Prävention von Rechenschwierigkeiten. In A. Fritz-Stratmann, G. Ricken & S. Schmidt (Hrsg.), *Handbuch Rechenschwäche* (3., überarb. Aufl., S. 339-355). Weinheim: Beltz.

Wielpütz, H. (2010). Qualtitätsanalyse und Lehrerbildung. In C. Böttinger, K. Bräuning, M. Nührenbörger, R. Schwarzkopf & E. Söbbeke (Hrsg.), *Mathematik im Denken der Kinder* (S. 109-114). Seelze: Klett-Kallmeyer.

Wilbert, J. (2014). Instrumente zur Lernverlaufsmessung: Gütekriterien und Auswertungsherausforderungen. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (Tests und Trends, N.F. Bd. 12, S. 281-308). Göttingen: Hogrefe.

Wilbert, J. & Linnemann, M. (2011). Kriterien zur Analyse eines Tests zur Lernverlaufsdiagnostik. *Empirische Sonderpädagogik, 3*, 225-242.

Winkelmann, H., Robitzsch, A., Stanat, P. & Köller, O. (2012). Mathematische Kompetenzen in der Grundschule. Struktur, Validierung und Zusammenspiel mit allgemeinen kognitiven Fähigkeiten. *Diagnostica, 58,* 15-30.

Zimmermann, K.R. (2011). *Jedes Kind kann rechnen lernen: Rechenschwäche und Dyskalkulie – Wie Eltern helfen können*. Weinheim: Beltz.

Dr. Simon Sikora
Universität Rostock
Insititut für Sonderpädagogische Entwicklungsförderung und Rehabilitation
August-Bebel-Straße 28
18055 Rostock
simon.sikora@uni-rostock.de

Erstmalig eingereicht: 10.03.2017
Überarbeitung eingereicht: 19.06.2017
Angenommen: 30.06.2017