Empirische Sonderpädagogik, 2017, Nr. 2, S. 165-183
ISSN 1869-4845 (Print) · ISSN 1869-4934 (Internet)

Theoriegeleitete Testkonstruktion dargestellt am Beispiel einer Lernverlaufsdiagnostik für den mathematischen Anfangsunterricht

Miriam Balt[1], Antje Ehlert[1] & Annemarie Fritz[2]

[1] *Universität Potsdam*
[2] *Universität Duisburg-Essen*

Zusammenfassung

Zur Erfassung individueller Lernentwicklungen in leistungsheterogenen Schulklassen werden aussagekräftige Verfahren zur Lernverlaufsdiagnostik benötigt, die adaptiv an die unterschiedlichen Lernausgangslagen der Kinder angepasst werden können. Die Entwicklung adaptiver Einzeltests kann nicht über parallele Messungen realisiert werden, sondern setzt eine alternative Herangehensweise an die Testkonstruktion voraus. Am Beispiel der Konstruktion einer Lernverlaufsdiagnostik für den mathematischen Anfangsunterricht wird im vorliegenden Beitrag die Vorgehensweise einer auf einem Entwicklungsmodell basierenden theoriegeleiteten Testentwicklung vorgestellt. Auf Basis des Entwicklungsmodells arithmetischer Konzepte (Fritz, Ehlert, & Balzer, 2013) wurden $N = 68$ Aufgaben konzipiert, welche die unterschiedlichen Entwicklungsniveaus des Modells operationalisieren. Diese Aufgaben wurden in einer längsschnittlichen Untersuchung mit $N = 279$ Erstklässler/innen einer empirischen Prüfung unterzogen und in Bezug auf ihre Änderungssensibilität untersucht. Ziel ist es, unter Verwendung der probabilistischen Testtheorie einen Aufgabenpool aufzubauen, der zukünftig auch für adaptives Testen eingesetzt werden kann. Die Aufgaben erwiesen sich als reliabel, valide und geschlechterfair und eignen sich zur Abbildung erster Lernentwicklungen. Es zeigte sich allerdings, dass die Aufgaben noch nicht alle Leistungsbereiche abdecken. Es bedarf weiterer schwierigerer Aufgaben, die die arithmetischen Konzepte der höheren Entwicklungsniveaus erfassen.

Schlagworte: Lernverlaufsdiagnostik, mathematischer Anfangsunterricht, adaptives Testen

A theory-based assessment of the learning process in primary school mathematics

Abstract

In order to assess individual learning progress in heterogeneous classrooms, sound progress monitoring measures are needed, which can be adjusted to the various levels of knowledge within a given class. The development of adaptive tests cannot be realized via parallel measurements and thus requires an alternative method of test construction. This article introduces the concept of a theory-driven test construction based on a developmental model, using the construction of a progress monitoring measure for early numeracy in primary school as an example. Based on the developmental model of arithmetic concepts (Fritz et al., 2013), $N = 68$ tasks were designed that operationalize the different developmental levels of the model. These tasks were empirical-

ly examined in a longitudinal study with $N = 279$ first grade students, focusing in particular on their responsiveness to learning progress. The purpose of this study is to generate an item pool using the item-response-theory, which can later be applied in adaptive tests. The tasks proved to be reliable, valid and gender fair, and are suitable for showing initial learning progress among students. However, it was found that the items do not cover all performance ranges. More difficult items are needed to measure the higher levels of the developmental model.

Keywords: progress monitoring, early numeracy measures, adaptive testing

Die deutsche Schulpraxis steht vor neuen Herausforderungen. Im Zuge der Umsetzung einer inklusiven Beschulung werden deutlich heterogenere Schülergruppen in den Grundschulen erwartet. Mit dem Recht der Schülerinnen und Schüler auf individuelle Unterstützung (z.B. Schulgesetz NRW, §1; UN-Behindertenrechtskonvention) verändern sich auch die Anforderungen, die an einen gemeinsamen Unterricht gestellt werden. Dazu gehört beispielsweise eine stärkere Berücksichtigung individueller Lernentwicklungen bei der Unterrichtsgestaltung und Planung von Fördermaßnahmen. Zur Erfassung dieser unterschiedlichen Lernentwicklungen werden aussagekräftige Verfahren zur Lernverlaufsdiagnostik benötigt. Lernverlaufsdiagnostische Verfahren versuchen mittels regelmäßig wiederholter Messungen die gegenwärtige Lernentwicklung abzubilden. Auf Basis dieser erhobenen Schülerdaten lassen sich pädagogische Förderentscheidungen treffen, welche eine individuelle Entwicklung berücksichtigen. Dieser Grundidee folgend werden in der Forschungsliteratur verschiedene Begrifflichkeiten wie „Progress Monitoring“ (z.B. Foegen, Jiban, & Deno, 2007; Salaschek & Souvignier, 2013, 2014) „Curriculum-based measurement (CBM)“ (z.B. Missall, Mercer, Martínez, & Casebeer, 2012), „Lernverlaufsdiagnostik“ (z.B. Klauer, 2011) oder „Lernfortschrittsmessung“ (Walter, 2008) weitgehend synonym verwendet.

In einer Vielzahl empirischer Studien konnte nachgewiesen werden, dass sich der Einsatz einer regelmäßigen Verlaufsdiagnostik positiv auf die Lernentwicklung von Schülerinnen und Schülern auswirkt (Kingston & Nash, 2011; Stecker, Fuchs, & Fuchs, 2005). Die alleinige Durchführung einer begleitenden Verlaufsdiagnostik ist dabei jedoch nicht hinreichend für den Lernerfolg. Entscheidenden Einfluss auf den Lernerfolg haben insbesondere Anpassungen des unterrichtlichen Handelns in Abhängigkeit der dokumentierten Lernentwicklung (Stecker et al., 2005) bzw. eine erfolgte effektive Leistungsrückmeldung an die Lernenden (Hattie & Timperley, 2007; Shute, 2008). Vor allem die Anpassung des unterrichtlichen Handelns stellt für die Lehrkräfte häufig eine besondere Herausforderung dar (Hojnoski, Gischlar, & Missall, 2009; Slavin, Cheung, Holmes, Madden, & Chamberlain, 2013). Dabei können qualitative Analysen der gelösten Aufgaben (also nicht nur wie viele, sondern auch welche Aufgaben gelöst wurden), die Verwendung datenbasierter Entscheidungsregeln und konkrete Handlungsempfehlungen eine wichtige Unterstützung bei der praktischen Umsetzung der Anpassung unterrichtlichen Handelns darstellen (Stecker et al., 2005).

Im Fokus des vorliegenden Beitrags steht die Konstruktion einer kompetenzorientierten Lernverlaufsdiagnostik für den mathematischen Anfangsunterricht. Diese soll in heterogenen Lerngruppen eingesetzt werden und es den Lehrkräften ermöglichen, adäquat auf die individuellen Lernentwicklungen ihrer Schülerinnen und Schüler im Unterricht einzugehen. Ausgehend von den vielfältigen Anforderungen an Verfahren zur Lernverlaufsdiagnostik werden grundlegende Herangehensweisen an die Testkonstruktion diskutiert. Anschließend wird das mathematische Kompetenzentwicklungsmodell beschrieben, welches die theoretische Grundlage für die Testkonstruktion bildet.

Zudem wird ein kurzer Überblick über aktuell verfügbare Verfahren zur Lernverlaufsdiagnostik von mengen- und zahlenbezogenen Basiskompetenzen gegeben und mögliche Limitationen der Verfahren für den Einsatz in heterogenen Lerngruppen aufgezeigt.

Anforderungen an eine Lernverlaufsdiagnostik

Die Anforderungen an eine Lernverlaufsdiagnostik, die sowohl aussagekräftig als auch praktikabel ist, sind vielfältig (z.B. Klauer, 2011; Wilbert & Linnemann, 2011). Souvignier, Förster und Schulte (2014) formulieren einen Anforderungskatalog auf drei Ebenen. Auf formaler Ebene sollen die Tests zeitökonomisch in Durchführung und Auswertung sein und eine unmittelbare Leistungsrückmeldung an die Schülerinnen und Schüler ermöglichen. Inhaltlich sollen die Tests so differenziert sein, dass die Anschlussfähigkeit für eine individuelle Förderung gewährleistet werden kann. Auf der Ebene der Testgüte müssen die klassischen Testgütekriterien (Objektivität, Reliabilität, Validität) erfüllt sein. Außerdem wird die Sensitivität der Einzeltests für Leistungsveränderungen gefordert. Um Lernentwicklung sichtbar zu machen, wird üblicherweise das Testkonzept der parallelen Messungen angewendet. Das bedeutet, dass die Einzeltests sowohl inhaltlich dasselbe erfassen, als auch Items gleicher Schwierigkeit beinhalten (Klauer, 2011). Die Entwicklung einer größeren Anzahl von Paralleltests, wie sie für eine Lernverlaufsdiagnostik benötigt wird, stellt häufig die größte Herausforderung bei der Testentwicklung dar (Souvignier et al., 2014). In Hinblick auf eine zunehmende Heterogenität der Schülerschaft wird außerdem der Ruf nach Testverfahren lauter, die unabhängig vom Lehrplan auch über Klassenstufen hinweg einsetzbar sind (Gebhardt, Heine, Zeuch, & Förster, 2015) und adaptiv an den Lernverlauf eines jeden Kindes angepasst werden können.

Um diesen Anforderungen und den unterschiedlichen diagnostischen Zielstellungen innerhalb des schulischen Lernens gerecht werden zu können, müssen bei der Konstruktion von Verfahren zur Lernverlaufsdiagnostik verschiedene grundlegende Entscheidungen getroffen werden.

Testtheorie: klassisch oder probabilistisch?

Klauer (2011) findet in Bezug auf die Wahl der Testtheorie klare Worte: „Die Verfahren der Lernverlaufsdiagnostik gemäß der klassischen Testtheorie zu konstruieren, ist praktisch ausgeschlossen“ (S. 211). Um den Anforderungen an eine Lernverlaufsdiagnostik aus testtheoretischer Sicht gerecht werden zu können, wird stattdessen die Verwendung der probabilistischen Testtheorie als „unumgänglich“ eingeschätzt (Wilbert & Linnemann, 2011, S.227). Als Hauptgrund für diese Einschätzung kann die mangelnde Überprüfbarkeit der Grundannahmen (z.B. homogener unkorrelierter Messfehler oder Intervallskalenniveau der gemessenen Skala) innerhalb der Klassischen Testtheorie angesehen werden. Eine Verletzung dieser Grundannahmen wird bei einmaliger Messung zur Statusdiagnostik als weniger kritisch eingeschätzt, fällt aber bei einer Veränderungsmessung im Kontext einer wiederholten Lernverlaufsdiagnostik deutlich stärker ins Gewicht (Wilbert, 2014). Die Spezifizierung eines gültigen Messmodells kann somit als die wichtigste testtheoretische Forderung an ein Instrument zur Lernverlaufsdiagnostik betrachtet werden, welche unter Verwendung der probabilistischen Testtheorie realisiert werden kann. Nach dieser Theorie wird eine latente Personeneigenschaft (Fähigkeit) über das Antwortverhalten einer Person auf entsprechende Items erfasst. Die gemessenen Werte sind demnach die mainfesten Indikatoren dieser latenten Fähigkeit, die es zu messen gilt. Das Antwortverhalten (z.B. das Lösen vs. Nichtlösen einer Aufgabe) hängt sowohl von Merkmalen der Person

(z.B. Fähigkeit, bestimmte mathematische Aufgaben zu lösen) als auch von Merkmalen der Aufgaben (z.B. deren Schwierigkeit) ab. Die Nutzung der probabilistischen Testtheorie zur Entwicklung von Lernverlaufsaufgaben sichert, dass die Personenfähigkeit nicht direkt mit den Rohwerten einer Testung gleichgesetzt (klassische Testtheorie), sondern durch ein mathematisches Modell (z.B. Rasch-Modell) spezifiziert wird, welches eine Wahrscheinlichkeit errechnet, ein spezifisches Item zu lösen. Eine Prüfung der Modellannahmen und der Items, inwiefern diese die latente Personeneigenschaft tatsächlich erfassen, kann durchgeführt werden. Für eine differenzierte Auseinandersetzung mit dem Rasch-Modell als ein valides Messmodell zur Veränderungsmessung im Kontext von Lernverlaufsdiagnostik sei hier auf Wilbert (2014) verwiesen.

Testkonstruktion: „curriculum sampling" oder „robust indicators"?

Bei der Konstruktion von Tests zur Lernverlaufsdiagnostik können nach Fuchs (2004) zwei grundsätzliche Herangehensweisen unterschieden werden - das „curriculum sampling" und die Verwendung von „robust indicators". Beim „curriculum sampling" werden Stichproben aus einem Pool an Aufgaben gezogen, die z.B. den Lehrplan eines bestimmten Schuljahres repräsentieren. Formative Testverfahren, die auf dieser Herangehensweise basieren, bestehen typischerweise aus parallelen Einzeltests. Diese müssen stets gleich schwer sein, sodass sich Lernentwicklungen über den Prozentsatz richtig gelöster Aufgaben erfassen lassen. Eine Entwicklung bildet sich somit dann ab, wenn im Laufe eines festgelegten Zeitintervalls der Prozentsatz richtig gelöster Aufgaben zunimmt. Vorteil des „curriculum samplings" ist die direkte Verbindung zum Lehrplan und damit die einfache Interpretierbarkeit der Testergebnisse für die Lehrkraft. Lernfortschritte können anschaulich visualisiert und für eine effektive Leistungsrückmeldung an die Lernenden eingesetzt werden. Qualitative Aussagen darüber, warum bestimmte Aufgaben nicht gelöst werden konnten, z.B. weil ein Kind noch nicht über hinreichende Einsichten in grundlegende arithmetische Konzepte verfügt, können in der Regel nicht getroffen werden. Hierfür bedarf es anderer Verfahren, die nicht curricular umschriebene Fähigkeiten erfassen, sondern auf Grundlage von Kompetenzentwicklungsmodellen konstruiert wurden.

Bei der Testkonstruktion auf Basis von „robust indicators" sollen Maße identifiziert werden, die nicht an den Lehrplan gebunden sind, sondern vielmehr spezifische Aspekte von Kernkompetenzen widerspiegeln (Foegen et al., 2007). Um einen solchen Test zu konstruieren und jene Kernkompetenzen sinnvoll zu operationalisieren, ist eine empirisch gesicherte Theorie über die zu messende Kompetenzentwicklung als Basis der Testentwicklung unerlässlich. Diese theoretische Fundierung ermöglicht es, die Aufgaben in einen inhaltlichen Zusammenhang zu bringen, z.B. in Bezug auf ihre zeitliche Abfolge im Entwicklungsprozess. Daraus ergeben sich konkrete Erwartungen an die Aufgabenschwierigkeiten. Das bedeutet, Aufgaben, die grundlegende Basiskompetenzen erfassen, werden leichter sein, als Aufgaben, die in einem hierarchischen Entwicklungsmodell darauf aufbauende Kompetenzen erheben. Diese Zusammenhänge lassen sich unter Verwendung der probabilistischen Testtheorie auch empirisch überprüfen. Damit scheint die Testkonstruktion mittels „robust indicators" die Methode der Wahl zu sein, wenn auf Basis des Testwerts auch qualitative Aussagen über den Entwicklungsstand der Kinder getroffen werden sollen. Aufgrund der Kompetenzorientierung kann von einer höheren inhaltlichen Förderrelevanz ausgegangen werden, indem der Lehrkraft, diejenigen Kompetenzbereiche aufgezeigt werden, an denen im Rahmen von Unterricht und Förderung weiter gearbeitet werden sollte (Voß, Sikora, & Hartke, 2017).

Einzeltests: parallel oder adaptiv?

Um Lernverläufe abbilden zu können, sollten nach Klauer (2011) die Einzeltests immer das gleiche Konstrukt erfassen und stets gleich schwer, also parallel sein. Bei lehrzielorientierten Verfahren ist es daher erforderlich, dass schon von Schuljahresbeginn die eigentliche Zielkompetenz des jeweiligen Schuljahres miterfasst wird. Das bedeutet, dass in den ersten Messungen den Kindern Aufgaben vorgelegt werden, welche sie mit ihrem derzeitigen Wissensstand noch nicht lösen können. Erst mit Voranschreiten des Schuljahres wird das Lösen dieser Aufgaben möglich. Nur so kann über die Anzahl der richtig gelösten Aufgaben ein Lernverlauf abgebildet und dieser in Verlaufskurven dargestellt werden (Klauer, 2011).

Kompetenzorientierte Aufgaben, die theoriegeleitet auf Basis eines Entwicklungsmodells konzipiert wurden, können hingegen auch in adaptiven Tests eingesetzt werden. Dabei lösen die Kinder eine Auswahl an Aufgaben, die auf Grundlage der Ergebnisse der vorherigen Messung an ihren individuellen Entwicklungsstand angepasst worden sind. Es liegen folglich keine parallelen Messungen vor, sondern die Schwierigkeit der Aufgaben ändert sich in Abhängigkeit von der qualitativen Lernentwicklung eines Kindes. In Hinblick auf eine leistungsheterogene Schülerschaft bzw. für einen jahrgangsübergreifenden Einsatz erscheint der Einsatz adaptiver Einzeltests im Vergleich zu parallelen Einzeltests besonders geeignet, da Aufgaben flexibel auf den Lernstand der Kinder angepasst werden können, ohne diese zu über- oder zu unterfordern.

Mathematische Kompetenzentwicklungsmodelle

Ziel jeder Lernverlaufsdiagnostik sollte die Anpassung des unterrichtlichen Handelns auf Basis der erhobenen Schülerdaten sein. Um die Lehrkräfte beim Treffen pädagogischer Förderentscheidungen zu unterstützen, sollten formative Diagnoseverfahren auf der Grundlage von Kompetenzentwicklungsmodellen konstruiert werden, sodass der Testwert nicht nur quantitativ im Vergleich zur Altersnorm interpretiert werden kann, sondern auch eine qualitative Aussage über den Entwicklungsstand liefert (Voß et al., 2017). Im Bereich der mathematischen Kompetenzentwicklung stehen bereits verschiedene Modelle zur Verfügung (Fritz et al., 2013; Krajewski & Schneider, 2009; Reiss & Winkelmann, 2008). Sie gehen von einem sukzessiven Kompetenzerwerb aus, bei dem aufeinander aufbauende Niveaus durchlaufen werden. Im Folgenden wird das Entwicklungsmodell arithmetischer Konzepte von Fritz, Ehlert und Balzer (2013) beschrieben, dessen Gültigkeit in unterschiedlichen Längsschnittstudien aufgezeigt werden konnte (Fritz, Ehlert, & Leutner, in Druck). Das Modell bildet die Grundlage der theoriegeleiteten Itemkonstruktion für die Lernverlaufsdiagnostik.

- Niveau I (Zählzahl): Die Kinder haben verstanden, dass Zahlen zum Zählen von Objekten eingesetzt werden können. Eine bedeutsame Strategie ist dabei die Eins-zu-Eins-Zuordnung. Sie können Mengen aus- und abzählen, ohne die Mächtigkeit der Menge (Kardinalität) zu kennen.
- Niveau II (ordinaler Zahlenstrahl): Es wird verstanden, dass die Zahlen in der Zahlwortreihe „größer" werden. Zählend können nun Additionen und Subtraktionen im kleinen Zahlenraum präzise ausgeführt und Vorgänger- und Nachfolgerzahlen in der Zahlwortreihe bestimmt werden.
- Niveau III (Kardinalität und Zerlegbarkeit): Zahlen werden mit der Mächtigkeit der entsprechenden Menge verbunden und die Zahlwortreihe als Sequenz steigender Mächtigkeit verstanden. Handelnd wird der Zusammenhang von Teilmenge – Teilmenge – Gesamtmenge erkannt.
- Niveau IV (Enthaltensein und Klasseninklusion): Die Inklusionsbeziehung der

Zahlen wird verstanden. Jede Zahl enthält die Menge der vorangegangenen Zahlen. Damit werden Zahlen in unterschiedliche Teilmengen zerlegbar. Der Zusammenhang von Teilmenge – Teilmenge – Gesamtmenge wird weiter elaboriert und kann nun über das Bestimmen der Gesamtmenge hinaus auch auf Textaufgaben mit Fragen nach der Austauschmenge angewendet werden. Im Anschluss daran kann die Startmenge bestimmt werden.

– Niveau V (Relationalität): Es wird verstanden, dass die Abstände zwischen zwei aufeinanderfolgenden Zahlen immer gleich groß (+1) sind. Entsprechend beschreibt ein Abschnitt auf dem Zahlenstrahl oder der Abschnitt zwischen zwei Zahlen einmal eine Menge, um die sich beide Angaben differenzieren und einmal die Anzahl der dazwischenliegenden Zahlen und damit eine Sequenz.
– Niveau VI (Bündeln und Entbündeln): Aufbauend auf den Einsichten in die flexible Zahlzerlegung und das Konzept der Relationalität beginnt ein Kind zu verstehen, dass Zahlen in Einheiten gleicher Größe zerlegt werden können (z.B. 6 in 3 und 3). Zahlen können als Vereinigung gleichmächtiger Teilmengen (Bündel) verstanden werden. Damit wird die Voraussetzung für das Verständnis der Multiplikation und Division und des Stellenwertsystems geschaffen.

Lernverlaufsdiagnostik früher mathematischer Kompetenzen

Die richtungsweisende Bedeutung der mengen- und zahlenbezogenen Basiskompetenzen für den Lernerfolg im Mathematikunterricht der späteren Grundschuljahre konnte bereits in einer Reihe von Studien aufgezeigt werden (Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Duncan et al., 2007, Krajewski & Schneider, 2009; Weißhaupt, Peucker, & Wirtz, 2006). Die Entwicklung und Erforschung geeigneter Testverfahren zur Lernverlaufsdiagnostik von mengen- und zahlenbezogenen Basiskompetenzen verzeichnete allerdings erst in den letzten Jahren einen nennenswerten Aufschwung. Gersten, Jordan und Flojo (2005) beschreiben das Forschungsfeld noch als in den Kinderschuhen steckend. Im englischen Sprachraum wurden vor allem Untersuchungen zu den sogenannten „TEN-CBMs“ (= Tests of Early Numeracy) mit Vorschulkindern bzw. mit Kindern der ersten Klasse durchgeführt (z.B. Baglici, Codding, & Tryon, 2010; Chard et al., 2005; Hampton et al., 2012; Lembke & Foegen, 2009). Die „TEN-CBMs“ werden nach dem Prinzip der „robust indicators“ konstruiert und dienen der Erfassung spezifischer Aspekte des sogenannten „number sense“ (z.B. Dehaene, 1997). Sie bestehen aus vier separaten Tests, welche die Fähigkeit der Kinder erheben, in einer Sequenz von drei Zahlen die fehlende Zahl zu erkennen („missing number“), Ziffern zwischen 0 und 20 zu benennen („number identification“), die größere Zahl eines Zahlenpaares ausfindig zu machen („quantity discrimination“) und die Zahlwortreihe von Eins beginnend aufzusagen („oral counting“). Auch wenn die Verfahren eine hohe Reliabilität im Bereich von r_{tt} = .78 und r_{tt} = .99 aufweisen (Foegen et al., 2007), wird vor allem ihre schwierige Interpretierbarkeit kritisiert, welche zu einer erschwerten Ableitung konkreter Fördermaßnahmen führt (Methe, 2012). Als eine mögliche Ursache kann die mangelnde Verbindung zum Lehrplan angenommen werden. Kritisch ist jedoch zu betonen, dass der Begriff „number sense“ nicht einheitlich definiert wird (Gersten et al., 2005). Den „TEN-CBMs“ wird daher vorgeworfen, dass sie wichtige konzeptuelle Aspekte des „number sense“, wie die Fähigkeit zur Eins-zu Eins- Zuordnung („touch counting“), das Verfügen über einen mentalen Zahlenstrahl („ordinality“), die Anwendung des Teil-Teil-Ganze Prinzips („decomposition“) oder das Konzept des Bündelns („group by five“) nicht berücksichtigen (Methe, Hojnoski, Begeny, & Leary, 2011). Dies könnte als Grund dafür gesehen werden, dass die Ef-

fektstärken der Vorhersage der „TEN-CBMs“ für spätere Leistungen im Mathematikunterricht stark variieren (Missall et al., 2012). Die Verwendung eines validen Entwicklungsmodells könnte hier Abhilfe schaffen und zur Verbesserung der Testgüte und der Interpretierbarkeit der Ergebnisse beitragen.

Im deutschsprachigen Raum erfuhr die Erfassung mathematischer Lernentwicklungen in der Grundschule in den letzten Jahren ebenfalls eine erhöhte Aufmerksamkeit, wenngleich die Anzahl der verfügbaren Testverfahren bislang noch recht überschaubar ist. Es wurden beispielsweise US-amerikanischen Verfahren, wie die obig beschriebenen „TEN-CBMs“, für das deutsche Schulsystem adaptiert und im Rahmen des Rügener Inklusionsmodells zur Lernverlaufsdiagnostik bereits ab der ersten Klasse eingesetzt (Voß, 2014). Auch die Tests des internetbasierten Testsystems „quop“ beinhalten Aufgaben, die mengen- und zahlenbezogene Basiskompetenzen erfassen und bereits im mathematischen Erstunterricht eingesetzt werden können (Souvignier, Förster, & Salaschek, 2010). Ab der zweiten Klasse liegt der Fokus der derzeitig verfügbaren Verfahren vor allem auf der Beherrschung der Grundrechenarten entsprechend des Rahmenlehrplans. Für diesen Altersbereich stehen weitere Verfahren wie die „Lernverlaufsdiagnostik – Mathematik für zweite bis vierte Klasse“ (LVD-M 2-4; Strathmann & Klauer, 2012) oder die „Lernfortschrittsdiagnostik: Grundrechenarten“ (Müller & Hartmann, 2014) zur Verfügung. Durch die Zuweisung der Tests zu einer spezifischen Klassenstufe und die Anbindung an den entsprechenden Lehrplan, spielen die mengen- und zahlenbezogenen Basiskompetenzen in den Tests für die höheren Klassenstufen nur noch eine geringe Rolle. Dies macht aus testökonomischer Sicht Sinn, wenn sich die Lernverlaufsdiagnostik aus parallelen Einzeltests zusammensetzt, da bereits sämtliche Fähigkeiten, die bis Ende des Schuljahres gekonnt werden sollen, von Anfang an getestet werden müssen. Die Einzeltests wären viel zu umfangreich und würden starke Bodeneffekte aufweisen, wenn zusätzlich zu den regulären Lehrplaninhalten die Basiskompetenzen erhoben werden würden.

Aus der kurzen Darstellung von Verfahren zur Lernverlaufsdiagnostik früher mathematischer Kompetenzen wird deutlich, dass die sehr wenigen verfügbaren Verfahren insbesondere im deutschen Sprachraum vor allem curricular ausgerichtet sind. Es lassen sich zudem bereits Grenzen der Verfahren erkennen, denen mit einer kompetenzorientierten Testentwicklung begegnet werden kann. Das betrifft zum einen die Probleme der Operationalisierung bestimmter Kernkompetenzen, so wie sie bei den „TEN-CBMs“ zu beobachten sind, zum anderen könnten die Barrieren zwischen den Klassenstufen, die durch eine Testkonstruktion nach dem Prinzip des „curriculum samplings“ entstehen, aufgehoben werden. Ziel zukünftiger Forschungsbemühungen sollte daher sein, die Bandbreite an diagnostischen Verfahren zu erhöhen, sodass je nach diagnostischer Zielstellung, wie etwa dem Erreichen eines Lehrziels oder der Entwicklung einer Kompetenz, ein geeignetes Verfahren von der Lehrkraft ausgewählt und eingesetzt werden kann.

Ziele und Forschungsfragen der Studie

Im Rahmen der Konstruktion einer kompetenzorientierten Lernverlaufsdiagnostik für den mathematischen Anfangsunterricht wurden auf Grundlage des Entwicklungsmodells arithmetischer Konzepte (Fritz et al., 2013) verschiedene Aufgaben konzipiert, die die zentralen arithmetischen Konzepte der Entwicklungsniveaus des Modells operationalisieren. In der vorliegenden Studie werden diese Aufgaben einer empirischen Prüfung unterzogen und ihre Sensibilität zur Erfassung von Lernentwicklungen untersucht. Ziel ist es, einen „nach den Prinzipien der probabilistischen Testtheorie kalibrierten Itempool“ (Gebhardt et al., 2015, S. 208) aufzubauen, der zukünftig für adap-

tives Testen eingesetzt werden kann. Dabei werden folgende Forschungsfragen untersucht:

1. Entsprechen die entwickelten Aufgaben den Testgütekriterien der Reliabilität, Validität und Testfairness?
2. Können mit diesen Aufgaben mathematische Leistungsveränderungen bereits in den ersten Monaten der ersten Klasse abgebildet werden?

Methode

Stichprobe / Durchführung

Es wurde eine längsschnittliche Erhebung der arithmetischen Leistungen von $N = 279$ Erstklässler/-innen aus drei Bundesländern durchgeführt. Rund die Hälfte der Kinder waren Mädchen (49 %). Die Rekrutierung der insgesamt acht teilnehmenden Grundschulen erfolgte auf Anfrage durch die Testleiterinnen und nach Interesse der Schulleitungen an einer Studienteilnahme. Die Erhebungen der Lernverläufe starteten im Schuljahr 2015/2016 circa fünf Wochen nach Einschulung mit einem circa sechswöchigen Abstand zwischen den drei Messzeitpunkten (t1 bis t3). Die Testungen erfolgten in Kleingruppen von maximal 15 Kindern und nahmen jeweils eine Schulstunde in Anspruch. Die Instruktionen wurden von der Testleiterin laut vorgelesen und die Kinder bearbeiteten die Aufgaben in einem Testheft. Die Testleiterinnen waren Mitarbeiterinnen des Lehrstuhls und geschulte Masterstudentinnen. Zur Validierung der eingesetzten Aufgaben bearbeitete eine Teilstichprobe von $N = 81$ Kindern an zwei der teilnehmenden Schulen zusätzlich ein standardisiertes mathematisches Testverfahren. Dieses wurde im Einzelsetting zeitlich vor dem ersten Messzeitpunkt der Lernverlaufsdiagnostik durchgeführt.

Testinstrumente

Itemkonstruktion

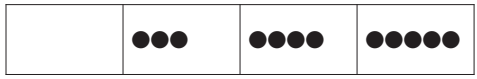
Bei der Itemkonstruktion und Weiterentwicklung von Aufgabenformaten sollten grundsätzlich zwei Phasen zirkulär durchlaufen werden, die theoretische Reflexion und die empirische Prüfung (Bühner, 2011). Im Rahmen der Itemkonstruktion für die Lernverlaufsdiagnostik wurden daher zunächst die zentralen arithmetischen Konzepte, damit verbundene Fähigkeiten (wie z.B. die Eins- zu- Eins- Zuordnung oder das Bestimmen von Vorgänger- und Nachfolgerzahlen) sowie deren hierarchische Anordnung im Laufe des Entwicklungsprozesses definiert. Dies erfolgte theoriegeleitet auf Basis des obig beschriebenen Entwicklungsmodells von Fritz, Ricken und Balzer (2013). Darauf aufbauend wurden entsprechende Items formuliert und im fachlichen Austausch zwischen den Projektmitgliedern diskutiert. Es entstanden neuartige Items, es wurden aber auch bereits vorhandene Aufgabenformate aus Lehrbüchern, Fördermaterialien und Tests wurden adaptiert. Die Items wurden anschließend einer empirischen Prüfung unterzogen. In einer unveröffentlichten Pilotierungsstudie mit $N = 447$ Erstklässler/ -innen wurden sie querschnittlich erprobt. Die Aufgabenformate wurden so konzipiert, dass sie in Gruppentestungen eingesetzt werden können. Tabelle 1 zeigt eine Auswahl an Aufgaben auf den unterschiedlichen Entwicklungsniveaus.

Erhebung des Lernverlaufs

Insgesamt wurden $N = 68$ arithmetische Aufgaben erprobt. Die Aufgaben wurden in einem Multi-Matrix-Design auf zwei Gruppen über drei Testheftversionen (A, B und C) mit je 30 Aufgaben verteilt (s. Tabelle 2).

Um eine möglichst große Anzahl an Aufgabenformaten prüfen zu können, wurde die Stichprobe in zwei Gruppen aufgeteilt. Gruppe 1 bearbeitete die Testheftkombination A-B-A und Gruppe 2 die Kombina-

Tabelle 1: Beispielaufgaben zur Erfassung der Entwicklungsniveaus

Niveau	Arithmetisches Konzept	Beispielaufgabe
I	Zählzahl	Malt genauso viele Punkte in das leere Kästchen, wie ihr hier seht. *(4 Punkte abgebildet)*
II	Ordinaler Zahlenstrahl	Wie heißt die Zahl die *vor* und wie heißt die Zahl die *nach* der 4 kommt? Schreibt die Zahlen in die leeren Felder.
III	Kardinalität	Hier seht ihr Kästchen mit Punkten. Aber in einem Kästchen fehlen Punkte. Wie viele Punkte müssen in das leere Kästchen? Malt so viele Punkte in das Kästchen bis die Reihe stimmt. 
IV	Enthaltensein und Klasseninklusion	Malt 5 Striche, davon sollen 3 rot sein.

tion C-B-C. Die Kinder innerhalb einer Gruppe lösten demnach zu t1 und t3 jeweils die gleichen Aufgaben (Testheftversion A oder C). Um Erinnerungseffekten aufgrund der Testwiederholung entgegen zu wirken, wurde allen Kindern zu t2 die Testheftversion B vorgelegt, die abgesehen von den 11 Ankeritems andere Aufgabenformate enthielt als die Testheftversionen A und C. Unter Ankeritems versteht man gemeinsame Aufgaben, die in allen Versionen (A, B und C) enthalten sind und entsprechend von jedem Kind zu jedem Messzeitpunkt bearbeitet werden. Sie bilden den Bezugspunkt zur Einschätzung der Schwierigkeiten der übrigen Items und erlauben damit eine gemeinsame Skalierung aller Aufgaben. Als Ankeritems wurden diejenigen Aufgaben ausgewählt, die in der Pilotierung die Entwicklungsniveaus am besten abbildeten (Prüfung über die Verteilung der Aufgaben auf der Raschskala) und bei denen keine Anpassungen, beispielsweise in Bezug auf die Aufgabenformulierung oder das Format notwendig waren. Die Anzahl der Aufgaben pro Testheft (N = 30) setzt sich wie folgt zusammen: 11 Ankeritems (zwei Aufgaben auf Niveau I und jeweils drei Aufgaben auf den Niveaus II, III und IV) und 19 Erprobungsaufgaben. Bei drei Testheftversionen ergibt das eine Gesamtanzahl von N = 68 Aufgaben.

Tabelle 2: Aufteilung der Testheftversionen über 3 Messzeitpunkte (t1 bis t3)

	Gruppe 1 (n = 166)	Gruppe 2 (n = 113)
t1	A ———	C
	|	|
t2	B ———	B
	|	|
t3	A ———	C

MARKO-D

Dieser standardisierte Test (Ricken, Fritz, & Balzer, 2013) ist ein Rasch-skaliertes Diagnoseverfahren, dem ebenfalls das oben erwähnte Entwicklungsmodell (Fritz et al., 2013) zugrunde liegt. Der Test wird im Gegensatz zur Lernverlaufsdiagnostik als Einzeltest durchgeführt. Aufgrund der unterschiedlichen Durchführungssettings liegt keine Überschneidung der Aufgabenformate zwischen den beiden Testverfahren vor. Der MARKO-D ist für den Altersbereich 4;0 bis 6;3 Jahre normiert und erfasst die Entwicklungsniveaus I bis V. Die Leistungen

können quantitativ (beruhend auf Roh- oder Personenfähigkeitswerten), aber auch qualitativ (beruhend auf der inhaltlichen Beschreibung im Rahmen des Entwicklungsmodells arithmetischer Konzepte) ausgewertet werden.

Statistische Auswertung

Passung der Aufgaben an das Raschmodell und an das Entwicklungsmodell arithmetischer Konzepte

Aufgrund der dichotomen Datenstruktur (richtig / falsch) wurde das einfache dichotome Raschmodell gewählt. Die Rasch-Modellpassung bestand aus einer quantitativen und einer qualitativen Überprüfung der Items. Im Rahmen der quantitativen Analyse wurde die Modellgültigkeit anhand von Prüfstatistiken untersucht, während die qualitative Analyse Auskunft über die inhaltliche Passung der Aufgaben zur Entwicklungstheorie gab. Als Prüfstatistiken standen sogenannte Infit- und Outfit- Werte zur Verfügung (Linacre, 2002). Beide Werte sind Maße für die Passung der Daten zum Modell und seinen Annahmen. Nach Linacre (2002) sind schlechte Outfit-Werte weniger bedeutsam für das Modell als auffällige Infit-Werte (siehe auch Adams & Wu, 2002). Mit der Infit-Statistik werden die tatsächlich beobachteten Lösungshäufigkeiten mit den auf Grundlage des angenommenen Modells berechneten Lösungswahrscheinlichkeiten basierend auf dem Gesamtantwortmuster im Datensatz verglichen.

Für die Niveaus des Entwicklungsmodells arithmetischer Konzepte (Fritz et al., 2013) wurden verschiedene Aufgaben formuliert, die das jeweilige arithmetische Konzept über unterschiedliche Aufgabenformate operationalisieren. Aufgrund der theoriegeleiteten Herangehensweise bestanden eindeutige Erwartungen an die Schwierigkeit eines Items und damit an die Position des Items innerhalb einer hierarchischen Raschskala. Es wurde angenommen, dass sich Testaufgaben, die auf der Basis gleicher Konzepte lösbar sind, auf einem Niveau gruppieren. Im Rahmen der qualitativen Analyse wurden diejenigen Aufgaben identifiziert, die diesen Erwartungen nicht entsprachen, also entweder schwerer oder leichter waren, als ursprünglich angenommen.

Die Verteilung der Personen und der Items auf der Raschskala gab weiterhin Auskunft über die Passung der Testschwierigkeit zur Kompetenzausprägung der Stichprobe. Waren Kinder oberhalb des schwierigsten Items zu finden, bedeutete dies, dass für einen Teil der Kinder die Testaufgaben eher zu leicht ausfielen. Dies traf auch zu, wenn im unteren Skalenbereich keine Personen zu finden waren, obgleich Items vorlagen. In diesem Fall waren die Testaufgaben demnach eher zu schwer.

Validität und Testfairness

Die Validität der Aufgaben des Lernverlaufs wurde über die konvergente Validität mit dem MARKO-D bestimmt. Dazu wurde die Korrelation zwischen den Personenfähigkeitswerten zum ersten Messzeitpunkt der Lernverlaufsdiagnostik und den Personenfähigkeitswerten des MARKO-D berechnet. Zudem wurde die geschlechtsspezifische Testfairness mittels „Differential Item Functioning“ (=DIF) ermittelt. Dadurch sollte sichergestellt werden, dass es zu keiner systematischen Benachteiligung eines der beiden Geschlechter beim Lösen der Aufgaben kam.

Sensibilität zur Erfassung von Lernentwicklungen

Zur Überprüfung, ob mit den eingesetzten Aufgaben bereits erste Lernentwicklungen über die drei Messzeitpunkte (t1 bis t3) sichtbar gemacht werden können, wurde mittels der Personenfähigkeitswerte als abhängige Variable eine Varianzanalyse mit Messwiederholung durchgeführt. In einem ersten Schritt wurden alle Kinder zusammen betrachtet. Anschließend wurden sie

in die drei Gruppen „schwach", „durchschnittlich" und „akzeleriert" aufgeteilt, um mögliche spezifische Entwicklungsverläufe in den unterschiedlichen Fähigkeitsbereichen aufzudecken. Die Gruppeneinteilung erfolgte auf Basis der gelösten Aufgaben auf den unterschiedlichen Entwicklungsniveaus. Dazu wurde in Anlehnung an die Auswertungsstandards des MARKO-D das 75%-Kriterium genutzt (Ricken et al., 2013). Das 75%-Kriterium besagt, dass eine Mindestanzahl an Aufgaben (75%) eines spezifischen Entwicklungsniveaus richtig bearbeitet werden muss, um das konzeptuelle Verständnis dieses Entwicklungsniveaus beim Kind als gesichert anzunehmen. Die Gruppe mit schwachen Leistungen (N = 36) umfasste folglich alle Kinder, die sich unter Anwendung des 75% Kriteriums zu t1 auf Niveau I und II befanden. Als durchschnittlich wurden die Kinder eingestuft (N = 167), die auf Niveau III und IV standen. Der Gruppe mit akzelerierten Leistungen (N = 68) wurden alle Kinder zugeordnet, die die entsprechende Mindestanzahl an Aufgaben auf Niveau IV richtig lösten.

Ergebnisse

Passung der Aufgaben an das Raschmodell und an das Entwicklungsmodell arithmetischer Konzepte

Bei der quantitativen Analyse der Modell-Passung wurden in einem dreischrittigen Vorgehen diejenigen Items aus dem Modell entfernt, deren Infit Werte nicht mindestens im akzeptablen Bereich (1 ± 0.3) lagen. Insgesamt ist dies bei drei Items der Fall. Die restlichen 65 Items weisen Infit-Werte im strengeren Bereich von 1 ± 0.2 bei einer Item- Reliabilität von .97 auf.

In der qualitativen Analyse wurde die Anordnung der Aufgaben innerhalb der Item-Person-Map geprüft. Aufgaben, die das gleiche arithmetische Konzept erheben, gruppieren sich auf der Raschskala und weisen folglich eine vergleichbare Schwierigkeit auf. 12 der 68 Aufgaben (18%) liegen nicht in den erwarteten Bereichen der Raschskala. Es wird geschlussfolgert, dass sie inhaltlich nicht das arithmetische Konzept messen, für dessen Erfassung sie konstruiert wurden. Aus diesem Grund wurden diese Aufgaben aus den weiteren Berechnungen ausgeschlossen. Um zu prüfen, wie die inhaltliche Nicht-Passung der Aufgaben zustande gekommen ist, wurden sie einer qualitativen Inhaltsanalyse unterzogen. Diese wird exemplarisch an einem Beispiel demonstriert: Vier bzw. sechs Ballons (visuell vorgegeben) sollen auf zwei Kinder aufgeteilt werden. Die jeweilige Anzahl pro Kind kann gemalt oder aufgeschrieben werden. Die Zerlegung von Mengen wird konzeptuell dem Niveau IV zugeordnet, allerdings konnten die Kinder aufgrund der Visualisierung und der damit möglichen Eins-zu-Eins-Zuordnung die Aufgabe mit geringeren Personenfähigkeitswerten und damit früher lösen, als auf dem Entwicklungsniveau IV. Dieser Aufgabentyp und somit beide zugehörigen Aufgaben waren folglich zu leicht.

Aus der Verteilung der Personen und der Items auf der Raschskala wird zudem deutlich, dass keine ideale Passung zwischen Testschwierigkeit und Kompetenzausprägung der Stichprobe vorliegt. Die schwierigste Aufgabe im Test weist eine Itemschwierigkeit von 3.22 auf, während es noch eine ganze Reihe an Kindern gibt, deren Fähigkeit mit Werten bis zu 5.78 über der höchsten Itemschwierigkeit liegt. Es fehlen folglich schwierigere Aufgaben, die die arithmetischen Kompetenzen im oberen Leistungsbereich differenzieren. Die erprobten Aufgaben sind für die Kinder der ersten Klasse eher zu leicht.

Validität und Testfairness

Die Korrelation zwischen den Personenfähigkeitswerten der Lernverlaufsdiagnostik zum ersten Messzeitpunkt und den Personenfähigkeitswerten des MARKO-D liegt bei $r = .70$.

Bezüglich der geschlechtsspezifischen Testfairness weisen sechs Items (nach Entfernung der Items, die sowohl der quantitativen als auch qualitativen Modellpassung nicht entsprechen) einen signifikanten DIF Contrast auf. Der DIF Contrast beschreibt die Differenz der Schwierigkeiten eines Items bei zwei verschiedenen Gruppen, im vorliegenden Fall zwischen Mädchen und Jungen. Nach Boone, Staver und Yale (2014) ist allerdings das alleinige Vorliegen eines signifikanten DIF Contrasts noch kein Nachweis eines bedeutungsvollen Unterschieds. Vielmehr sagt erst die Effektgröße etwas über die Relevanz eines DIF's aus. Ein DIF Contrast $> |.64|$ deutet auf einen großen und damit bedeutsamen Unterschied zwischen der Leistungsmessung eines Items in zwei verschiedenen Gruppen hin. Fünf Items weisen einen bedeutsamen DIF auf. Diese Items messen die arithmetische Kompetenz der Mädchen bzw. der Jungen somit unterschiedlich. Die Items verteilen sich wie folgt: Zwei Items auf Level II, zwei Items auf Level III und ein Item auf Level IV. Es zeigt sich weiterhin, dass die Richtung des DIF's zwischen den Geschlechtern variiert und nicht systematisch ist. Zwei der fünf Items sind für Mädchen leichter zu lösen als für Jungen. Jungen haben bei drei Items einen Leistungsvorteil gegenüber den Mädchen. Es wird folglich angenommen, dass kein Geschlecht systematisch benachteiligt wird, weshalb die Items mit einem DIF im Aufgabenpool verbleiben. Bezogen auf die Gesamtanzahl der eingesetzten Items weisen 91% der Items keinen DIF auf.

Sensibilität zur Erfassung von Lernentwicklungen

Überprüfung über die Personenfähigkeitswerte

Die Varianzanalyse mit Messwiederholung zeigt, dass die durchschnittlichen Personenfähigkeitswerte aller Kinder von t1 zu t3 ansteigen. Es kann sowohl ein signifikanter Unterschied zwischen den drei Messzeitpunkten, $F(2, 245) = 58{,}176$, $p < .001$, als auch signifikante Kontraste zwischen t1 und t2, $F(1{,}246) = 29{,}563$, $p < .001$, sowie zwischen t2 und t3, $F(1{,}246) = 64{,}510$, $p < .001$, festgestellt werden (s. Abbildung 1).

Der Gruppenvergleich (schwach, durchschnittlich, akzeleriert) verdeutlicht, dass die Kinder auf sehr unterschiedlichen Fähigkeitsniveaus in den arithmetischen Anfangsunterricht der Grundschule starten (s. Tabelle 3).

In der akzelerierten Gruppe kann der weitere Lernverlauf über die Messzeitpunkte t2 und t3 allerdings auf Grund des Fehlens schwierigerer Aufgaben, welche die arithmetischen Kompetenzen im oberen

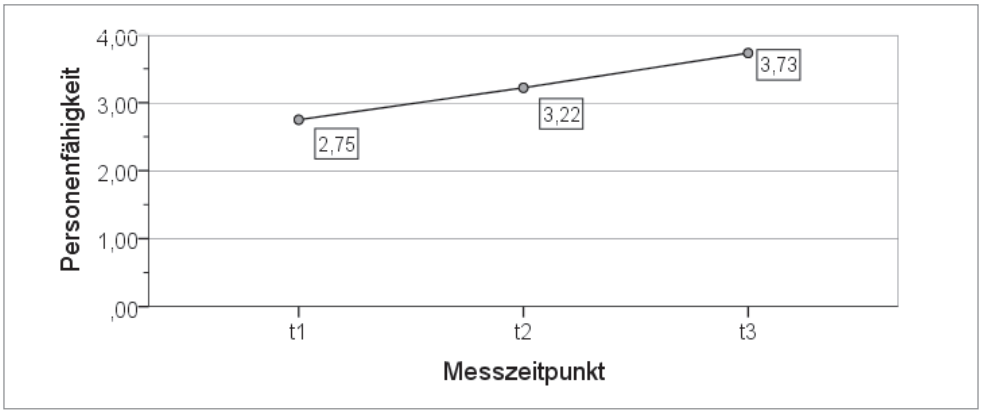


Abbildung 1: Leistungsentwicklung über 3 Messzeitpunkte (t1 bis t3)

Tabelle 3: Mittelwerte und Standardabweichungen der Personenfähigkeitswerte über 3 Messzeitpunkte (t1 bis t3) in Abhängigkeit von der Gruppenzugehörigkeit

	Schwach (n = 36)		Durchschnittlich (n = 167)		Akzeleriert (n = 68)	
	M	SD	M	SD	M	SD
t1	1.16	1.02	2.35	1.08	4.63	1.09
t2	1.84	1.46	3.03	1.39	4.46	1.06
t3	1.94	1.61	3.63	1.51	4.85	1.11

Leistungsbereich differenzieren, nicht sinnvoll interpretiert werden (vgl. Ergebnisse der Rasch-Analyse). Aus diesem Grund wird in Abbildung 2, welche die Leistungsentwicklung über die drei Messzeitpunkte in Abhängigkeit von der Gruppenzugehörigkeit zeigt, auf die Darstellung des Lernverlaufs der akzelerierten Gruppe verzichtet.

In der Gruppe der Kinder mit durchschnittlichen Leistungen liegt ein signifikanter Anstieg der Personenfähigkeitswerte über die drei Messzeitpunkte vor, $F(2,148) = 69.81$, $p < .001$, mit signifikanten Kontrasten zwischen t1 und t2, $F(1,149) = 36.45$, $p < .001$, sowie zwischen t2 und t3, $F(1,149) = 64.39$, $p < .001$. Auch in der Gruppe der leistungsschwachen Kinder unterscheiden sich die Personenfähigkeitswerte zu den drei Messzeitpunkten signifikant voneinander, $F(2,33) = 5.40$, $p = .009$, mit einem signifikanten Kontrast zwischen t1 und t2, $F(1,34) = 9.02$, $p = .005$. Die Veränderung der Personenfähigkeitswerte von t2 zu t3 ist in dieser Gruppe allerdings nicht mehr signifikant, $F(1,34) = 0.09$, $p = .765$ (s. Abbildung 2).

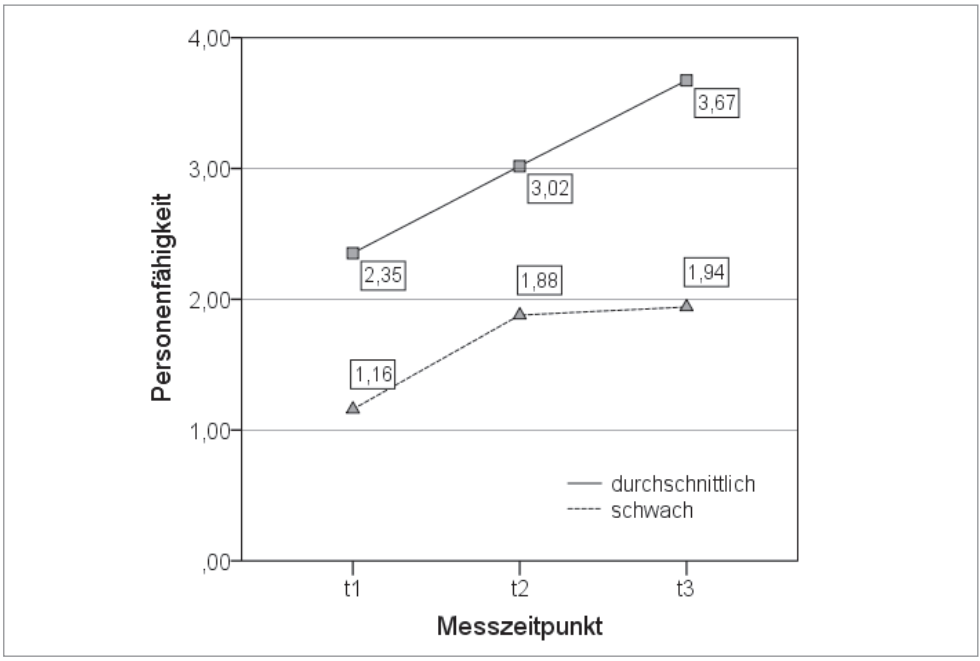


Abbildung 2: Leistungsentwicklung über 3 Messzeitpunkte (t1 bis t3) in Abhängigkeit von der Gruppenzugehörigkeit

Überprüfung über die Lösungsmuster

Nach quantitativer und qualitativer Analyse konnten diejenigen Aufgaben sowohl empirisch als auch inhaltlich abgesichert werden, die die zentralen arithmetischen Konzepte der einzelnen Entwicklungsniveaus verlässlich operationalisieren. Mit diesen Aufgaben erfolgte eine Analyse der Lösungsmuster, um zu prüfen, ob Lernentwicklungen auch über diese Lösungsmuster abgebildet werden können. Dazu wurden pro Messzeitpunkt die Summen der richtig gelösten Aufgaben für jedes Entwicklungsniveau gebildet. Um zu entscheiden, auf welchem Entwicklungsniveau ein Kind steht, wurde das obig beschriebene 75% Kriterium angewendet. Abbildung 3 zeigt die Verteilung der Kinder auf die verschiedenen Entwicklungsniveaus über die drei Messzeitpunkte (t1 bis t3) und damit die Entwicklung der arithmetischen Konzepte bzw. den Lernzuwachs der Kinder. Es wird deutlich, dass die Anzahl der Kinder auf den niedrigeren Entwicklungsniveaus über die Zeit abnimmt, die Anzahl der Kinder auf den höheren Entwicklungsniveaus hingegen steigt. Es sei an dieser Stelle angemerkt, dass aufgrund der Itemreduzierung nach vorangegangener quantitativer und qualitativer Analyse, die Entwicklungsniveaus nicht mehr durch die gleiche Anzahl an Items erhoben wurden. Es wird aber davon ausgegangen, dass mit Verwendung des 75%-Kriteriums trotzdem eine relative Vergleichbarkeit der einzelnen Niveaus geschaffen werden konnte.

Diskussion

Ziel des vorliegenden Beitrags ist es, das Konzept der theoriegeleiteten Konstruktion eines kompetenzorientierten Testverfahrens zur Lernverlaufsdiagnostik vorzustellen. Damit soll eine Form der Testkonstruktion aufgezeigt werden, die es ermöglicht, Verfahren zu entwickeln, die auch den Anforderungen in leistungsheterogenen Lerngruppen gerecht werden. In der dargestellten

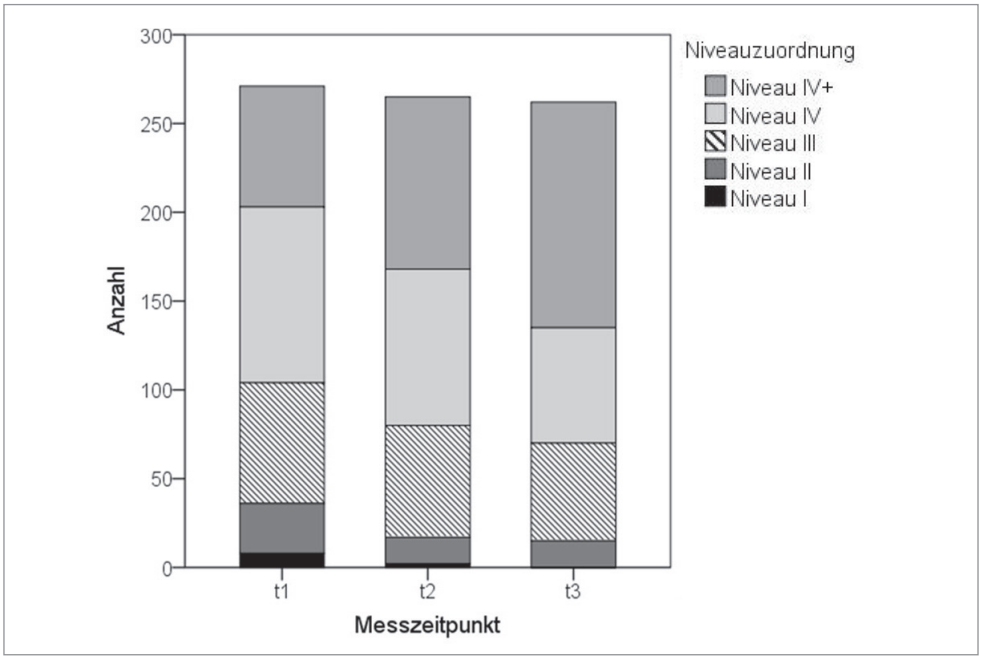


Abbildung 3: Verteilungen der Kinder auf die Entwicklungsniveaus zu den 3 Messzeitpunkten (t1 bis t3)

Studie wurden Aufgaben, die basierend auf dem Entwicklungsmodell arithmetischer Konzepte (Fritz et al., 2013) konzipiert wurden, einer empirischen Prüfung unterzogen, mit dem Ziel, einen Aufgabenpool aufzubauen, der zukünftig auch für adaptives Testen eingesetzt werden kann. Dazu wurden zwei zentrale Forschungsfragen untersucht.

Als erstes wurde geprüft, ob die entwickelten Aufgaben den Testgütekriterien der Reliabilität, Validität und Testfairness entsprechen. Nach Analyse der Modellpassung konnte festgestellt werden, dass 78% der erprobten Aufgaben die quantitativen und qualitativen Kriterien erfüllen und damit im Aufgabenpool verbleiben. Diese Aufgaben weisen eine hohe Itemreliabilität auf und können zudem als (konvergent) valide angesehen werden. Bezogen auf die Testfairness zeigen nur wenige Aufgaben einen DIF, dessen Richtung zudem variiert, sodass kein Geschlecht systematisch benachteiligt wird. Aus diesem Grund werden die Items mit einem DIF aktuell nicht aus dem Aufgabenpool entfernt und es wird angenommen, dass insgesamt geschlechterfaire Aufgaben vorliegen. Die Aufgaben decken allerdings noch nicht alle Leistungsbereiche ab. Es fehlen schwierigere Aufgaben. Als mögliche Ursache für die Unterschätzung der Schülerleistungen bei der Planung des Testmaterials in der vorliegenden Studie und damit der Verwendung von zu leichten Aufgaben, werden Stichprobeneffekte aus der vorrangegangen Pilotierungsstudie gesehen. Hier zeigten die Schülerinnen und Schüler schwächere Leistungen, die mit Aufgaben bis Niveau IV ohne Deckeneffekte erfasst werden konnten. Es wurde jedoch deutlich, dass bereits am Anfang der ersten Klasse auch Aufgaben benötigt werden, die die arithmetischen Konzepte der Entwicklungsniveaus V und VI erfassen. Aus diesem Grund werden in Folgestudien auch schwierigere Aufgaben erprobt, um den vorliegende Aufgabenpool zu ergänzen.

Als zweites wurde der Frage nachgegangen, ob mit den untersuchten Aufgaben bereits mathematische Leistungsveränderungen in den ersten Monaten der ersten Klasse abgebildet werden können. Es zeigte sich, dass die Aufgaben änderungssensibel sind und Entwicklungen im Abstand von sechs Wochen abbilden. Die Sensibilität wurde sowohl über die gesamte Stichprobe als auch für Kinder mit durchschnittlicher oder schwacher Leistung sichtbar. Die Entwicklung konnte anhand von Entwicklungskurven dargestellt werden, aber auch in Form von sich verändernden Verteilungen der Kinder auf den Entwicklungsniveaus. In beiden Darstellungsformen wurde deutlich, dass sich mit fortlaufender Dokumentation der konzeptuellen Entwicklung, weniger Kinder auf unteren Entwicklungsniveaus befinden und der Anteil der Kinder auf höheren Niveaus steigt. Inwiefern dies auch im höheren Kompetenzbereich möglich ist, muss in weiteren Untersuchungen mit schwierigeren Aufgaben geprüft werden.

Bevor abschließend ein Ausblick auf die konkrete Umsetzung des geplanten Verfahrens gegeben wird, soll noch kurz auf weitere Limitationen der vorliegenden Studie hingewiesen werden. So entspricht beispielsweise eine Testdauer von 45 Minuten noch nicht der Vorstellung eines ökonomischen Kurztests, der leicht in den Unterrichtsalltag integriert werden kann. Außerdem erfolgte im aktuellen Design noch keine Anpassung der Tests auf die individuelle Lernausgangslage der Kinder. Zudem kann auch die Anzahl von lediglich drei Messzeitpunkten kritisch gesehen werden.

In einer Folgestudie wird daher die arithmetische Lernentwicklung von Schülerinnen und Schülern der ersten Klasse über neun Messzeitpunkte anhand adaptiver Kurztests mit einer Dauer von maximal 15 Minuten erhoben. Zusätzlich findet eine differenzierte Eingangs- und Abschlussdiagnostik statt, in der neben der Bestimmung des arithmetischen Konzeptverständnisses auch Tests zu sprachlichen Fähigkeiten und der kognitiven Leistungsfähigkeit, sowie ein zusätzlicher Mathematiktest durchgeführt werden.

Ausblick: Umsetzung des geplanten Diagnoseverfahrens

Um die Lernverlaufsdiagnostik auf den individuellen Lernstand der Schülerinnen und Schüler anpassen zu können, muss zunächst ihr arithmetisches Konzeptverständnis im Vorfeld möglichst genau bestimmt werden. Steht ein Kind beispielsweise eingangs auf Niveau III, bekommt es im Rahmen der Lernverlaufsdiagnostik sowohl Aufgaben auf diesem Niveau, als auch den beiden angrenzenden Niveaus II und IV präsentiert. Die Auswahl der Aufgaben soll computergestützt aus dem Pool empirisch gesicherter Items erfolgen, sodass randomisiert immer neue Testaufgaben zusammengestellt werden können, die jeweils die entsprechenden Niveaus erheben. Eine Lehrkraft kann somit zu jedem Messzeitpunkt automatisiert Testhefte mit unterschiedlichen Aufgabenformaten generieren, die das derzeitige Entwicklungsfenster eines Kindes erfassen. Es liegen keine parallelen Messungen vor, da sich die Schwierigkeit der Testhefte in Abhängigkeit von der qualitativen Lernentwicklung eines Kindes ändert. Die Lernentwicklung wird nicht über die Gesamtanzahl richtig gelöster Aufgaben abgebildet, sondern darüber, welche Aufgaben vom Kind richtig bearbeitet wurden und welche nicht. Wird ein gewisser Prozentsatz an Aufgaben, die eine bestimmte Kernkompetenz erfassen, korrekt gelöst, kann diese Kompetenz als erworben angesehen werden. In diesem Fall werden dem Kind bei der nächsten Messung Aufgaben auf dem nächst höheren Entwicklungsniveau sowie Aufgaben auf den beiden benachbarten Niveaus präsentiert. Der erfasste Entwicklungsausschnitt verändert sich folglich und damit auch die Schwierigkeit der Testhefte.

Die Durchführung der Tests soll in Papierform erfolgen, da dafür keine spezielle technische Ausstattung an den Schulen erforderlich ist und sich die Diagnostik einfacher in den Unterrichtsalltag integrieren lässt. Die Tests werden von der Lehrkraft instruiert, können im Einzelsetting oder in Kleingruppen durchführt werden und sollen eine maximale Durchführungszeit von 15 Minuten nicht überschreiten. Die Auswertung der Testergebnisse kann von der Lehrkraft computergestützt durchgeführt werden, indem sie die Testergebnisse in ein Programm einträgt. Die Zuordnung und Erstellung des Testheftes für die nächste Messung erfolgt automatisiert. Zudem sollen individuelle Entwicklungsverläufe auf Grundlage der Niveauzuordnungen in Form von Lernkurven visualisiert werden können. Das qualitative Abbild der Lernentwicklung soll über die Lösungsmuster in einem Ampelsystem erfolgen. Hier kann pro Messzeitpunkt dargestellt werden, welches arithmetische Konzept schon sicher beherrscht wird (grün), welches Konzept vom Kind gerade erschlossen wird (gelb) und welches Verständnis ein Kind noch nicht haben kann, weil es in seiner Entwicklung die grundlegenden Kompetenzen noch nicht erworben hat (rot). Dieses System vereinfacht es den Lehrkräften zu erkennen, an welcher Stelle (gelb) ihr Unterricht bzw. eine zusätzliche Förderungen ansetzen muss, um ein Kind bestmöglich zu unterstützen. Es ist angedacht, in dem Auswertungsprogramm bereits konkrete Handlungsempfehlungen zu implementieren. Aufgrund der Einfachheit der Ampelanalogie kann diese Form der Visualisierung der Testergebnisse auch als direktes Feedback für die Lernenden eingesetzt werden.

Die vorliegende Studie lässt sich somit als erster Schritt verstehen, ein solches kompetenzorientiertes Verfahren zur Lernverlaufsdiagnostik für den mathematischen Anfangsunterricht zu entwickeln. Ziel ist es, die Lehrkräfte im Treffen pädagogischer Förderentscheidungen auf Basis der erhobenen Schülerdaten zu unterstützen, um auch in heterogenen Lerngruppen allen Kindern einen bestmöglichen Lernerfolg zu gewährleisten.

Literatur

Adams, R. J., & Wu, M. (2002). *PISA 2000 technical report*. Paris: OECD.

Aunola, K., Leskinen, E., Lerkkanen, M.-K., & Nurmi, J.-E. (2004). Developmental Dynamics of Math Performance From Preschool to Grade 2. *Journal of Educational Psychology*, *96*(4), 699–713. https://doi.org/10.1037/0022-0663.96.4.699

Baglici, S. P., Codding, R., & Tryon, G. (2010). Extending the Research on the Tests of Early Numeracy: Longitudinal Analyses Over Two School Years. *Assessment for Effective Intervention*, *35*(2), 89–102. https://doi.org/10.1177/1534508409346053

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Dordrecht: Springer.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. aktualisierte). München: Pearson-Studium.

Chard, D. J., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using Measures of Number Sense to Screen for Difficulties in Mathematics: Preliminary Findings. *Assessment for Effective Intervention*, *30*(2), 3–14.

Dehaene, S. (1997). *The Number Sense: How the mind creates mathematics*: Oxford University Press.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P.,. . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Foegen, A., Jiban, C., & Deno, S. (2007). Progress Monitoring Measures in Mathematics: A Review of the Literature. *The Journal of Special Education*, *41*(2), 121–139. https://doi.org/10.1177/00224669070410020101

Fritz, A., Ehlert, A., & Balzer, L. (2013). Development of mathematical concepts as basis for an elaborated mathematical understanding. *South African Journal of Childhood Education*, *3*(1), 38–67.

Fritz, A., Ehlert, A., & Leutner, D. (in Druck). Arithmetische Konzepte aus kognitiv-entwicklungspsychologischer Sicht. *Journal für Mathematik-Didaktik*.

Fuchs, L. S. (2004). The Past, Present, and Future of Curriculum-based Measurement Research. *School Psychology Review*, *33*(2), 188–192.

Gebhardt, M., Heine, J.-H., Zeuch, N., & Förster, N. (2015). Lernverlaufsdiagnostik im Mathematikunterricht der zweiten Klasse: Raschanalysen und Empfehlungen zur Adaption eines Testverfahrens für den Einsatz in inklusiven Klassen. *Empirische Sonderpädagogik*. (3), 206–222.

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early Identification and Interventions for Students With Mathematics Difficulties. *Journal of Learning Disabilities*, *38*(4), 293–304. https://doi.org/10.1177/00222194050380040301

Hampton, D. D., Lembke, E. S., Lee, Y.-S., Pappas, S., Chiong, C., & Ginsburg, H. P. (2012). Technical Adequacy of Early Numeracy Curriculum-based Progress Monitoring Measures for Kindergarten and First-Grade Students. *Assessment for Effective Intervention*, *37*(2), 118–126. https://doi.org/10.1177/1534508411414151

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hojnoski, R. L., Gischlar, K. L., & Missall, K. N. (2009). Improving Child Outcomes With Data-Based Decision Making: Graphing Data. *Young Exceptional Children*, *12*(4), 15–30. https://doi.org/10.1177/1096250609337696

Kingston, N., & Nash, B. (2011). Formative Assessment: A Meta-Analysis and a Call for Research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37.

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

Krajewski, K., & Schneider, W. (2009). Early development of quantity to number-word

linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning and Instruction*, *19*(6), 513–526. https://doi.org/10.1016/j.learninstruc.2008.10.002

Lembke, E., & Foegen, A. (2009). Identifying Early Numeracy Indicators for Kindergarten and First-Grade Students. *Learning Disabilities Research & Practice*, *24*(1), 12–20. https://doi.org/10.1111/j.1540-5826.2008.01273.x

Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Methe, S. A. (2012). Innovations and Future Directions for Early Numeracy Curriculum-based Measurement. *Assessment for Effective Intervention*, *37*(2), 67–69. https://doi.org/10.1177/1534508411431256

Methe, S. A., Hojnoski, R., Begeny, J. C., & Leary, L. L. (2011). Development of Conceptually Focused Early Numeracy Skill Indicators. *Assessment for Effective Intervention*, *36*(4), 230–242. https://doi.org/10.1177/1534508411414150

Missall, K. N., Mercer, S. H., Martínez, R. S., & Casebeer, D. (2012). Concurrent and Longitudinal Patterns and Trends in Performance on Early Numeracy Curriculum-based Measures in Kindergarten Through Third Grade. *Assessment for Effective Intervention*, *37*(2), 95–106. https://doi.org/10.1177/1534508411430322

Müller, C. M., & Hartmann, E. (2014). *Lernfortschrittsdiagnostik Grundrechenarten: 120 Drei-Minuten Tests für den inklusiven Mathematikunterricht - ZR bis 100*. Hamburg: Persen-Verlag.

Reiss, K., & Winkelmann, H. (2008). Step by step. Ein Kompetenzstufenmodell für das Fach Mathematik. *Grundschule*. (10), 34–37.

Ricken, G., Fritz, A., & Balzer, L. (2013). *MARKO-D. Mathematik- und Rechenkonzepte im Vorschulalter - Diagnose*. Göttingen: Hogrefe.

Salaschek, M., & Souvignier, E. (2013). Web-based progress monitoring in first grade mathematics. *Frontline Learning Research*, *1*(2). https://doi.org/10.14786/flr.v1i2.51

Salaschek, M., & Souvignier, E. (2014). Web-Based Mathematics Progress Monitoring in Second Grade. *Journal of Psychoeducational Assessment*, *32*(8), 710–724.

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Slavin, R. E., Cheung, A., Holmes, G., Madden, N. A., & Chamberlain, A. (2013). Effects of a Data-Driven District Reform Model on State Assessment Outcomes. *American Educational Research Journal*, *50*(2), 371–396. https://doi.org/10.3102/0002831212466909

Souvignier, E., Förster, N., & Salaschek, M. (2010). Ergebnisbericht des „quop"-Forschungsprojekts zur Lernverlaufsdiagnostik für das Schuljahr 2009/10. Retrieved from https://www.quop.de/fileadmin/literatur/Souvignier_Wissenschaftlicher_Ergebnisbericht_quop_2010.pdf

Souvignier, E., Förster, N., & Schulte, E. (2014). Wirksamkeit formativen Assessments - Evaluation des Ansatzes der Lernverlaufsdiagnostik. In M. Hasselhorn, W. Schneider, & U. Trautwein (Eds.), *Jahrbuch der pädagogisch-psychologischen Diagnostik. Tests und Trends: Band 12. Lernverlaufsdiagnostik* (pp. 221–237). Göttingen: Hogrefe.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using Curriculum-based Measurement to Improve Student Achievement: Review of Research. *Psychology in the Schools*, *42*(8), 795–819. https://doi.org/10.1002/pits.20113

Strathmann, A. M., & Klauer, K. J. (2012). *LVD-M 2-4. Lernverlaufsdiagnostik Mathematik für die zweiten bis vierten Klassen*. Göttingen: Hogrefe.

Voß, S. (2014). *Curriclumbasierte Messverfahren im mathematischen Erstunterricht: Zur Güte und Anwendbarkeit einer Adap-*

tion US-amerikanischer Verfahren im deutschen Schulsystem. Saarbrücken: SVH.

Voß, S., Sikora, S., & Hartke, B. (2017). Lernverlaufsdiagnostik als zentrales Element der Prävention von Rechenschwäche. In A. Fritz, S. Schmidt, & G. Ricken (Eds.), *Handbuch Rechenschwäche* (3rd ed., pp. 339–355). Beltz.

Walter, J. (2008). Curriculumbasiertes Messen (CBM) als lernprozessbegleitende Diagnostik: Erste deutschsprachige Ergebnisse zur Validität. Reliabilität und Veränderungssensibilität eines robusten Indikators zur Lernfortschrittsmessung beim Lesen. *Heilpädagogische Forschung*, *34*, 62–79.

Weißhaupt, S., Peucker, S., & Wirtz, M. (2006). Diagnose mathematischen Vorwissens im Vorschulalter und Vorhersage von Rechenleistungen und Rechenschwierigkeiten in der Grundschule. *Psychologie in Erzeihung und Unterricht*, *53*, 236–245.

Wilbert, J. (2014). Instrumente zur Lernverlaufsmessung: Gütekriterien und Auswertungsherausforderungen. In M. Hasselhorn, W. Schneider, & U. Trautwein (Eds.), *Jahrbuch der pädagogisch-psychologischen Diagnostik. Tests und Trends: Band 12. Lernverlaufsdiagnostik* (pp. 281–308). Göttingen: Hogrefe.

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

Miriam Balt
Universität Potsdam
Humanwissenschaftliche Fakultät
Inklusionspädagogik
Karl-Liebknecht-Straße 24-25
14476 Potsdam
miriam.balt@uni-potsdam.de

Erstmalig eingereicht: 07.03.2017
Überarbeitung eingereicht: 11.06.2017
Angenommen: 31.07.2017