Schwenk, Christin; Kuhn, Jörg-Tobias; Doebler, Philipp; Holling, Heinz

Auf Goldmünzenjagd: Psychometrische Kennwerte verschiedener Scoringansätze bei computergestützter Lernverlaufsdiagnostik im Bereich Mathematik

Empirische Sonderpädagogik 9 (2017) 2, S. 123-142

Empirische Sonderpädagogik, 2017, Nr. 2, S. 123-142
ISSN 1869-4845 (Print) · ISSN 1869-4934 (Internet)

Auf Goldmünzenjagd: Psychometrische Kennwerte verschiedener Scoringansätze bei computergestützter Lernverlaufsdiagnostik im Bereich Mathematik

Christin Schwenk[1], Jörg-Tobias Kuhn[1], Daniela Gühne[2], Philipp Doebler[2] & Heinz Holling[1]

[1] *Westfälische Wilhelms-Universität Münster*
[2] *Technische Universität Dortmund*

Zusammenfassung

In diesem Beitrag wird der computergestützte Lernverlaufstest „Goldmünzenjagd" vorgestellt, der in ein Online-Training für Kinder mit Rechenschwierigkeiten eingebettet ist. Der nach dem *robust indicator*-Ansatz konstruierte Test bildet den Lernfortschritt in zwei wichtigen mathematischen Basiskompetenzen ab: dem arithmetischen Faktenwissen (Addition bzw. Subtraktion bis 20) und dem Zahlenordnen (Zahlenreihen mit drei Elementen bis 100). Mit einem *High speed, high stakes*-Scoring wird die Bearbeitungseffizienz bewertet. Dieses Scoring verknüpft Geschwindigkeit und Präzision zu einem Gewinn oder Verlust von Goldmünzen auf Itemebene und zeigte sich in einer Feldstudie mit $N = 241$ Grundschulkindern (Klassenstufe 2 bis 4) sowohl in der Reliabilität ($r = .87$-$.93$) als auch in der Kriteriumsvalidität ($r = .51$) den klassischen Geschwindigkeits- und Präzisions-Scorings überlegen. Die individuellen Ergebnisse in den Lernverlaufstests waren zudem änderungssensitiv für die statusdiagnostische Entwicklung der Kinder: Für alle drei untersuchten Scorings ergab sich eine inkrementelle Varianzaufklärung der Leistung nach dem Training durch Parameter individueller Lernverläufe (*random intercept*: Ausgangspunkt Lernverlaufstest, *random slope*: Zuwachs Lernverlaufstest). Der vorgestellte Lernverlaufstest eignet sich damit als reliables und valides Tool zur formativen Evaluation der Leistungsentwicklung von Grundschulkindern in basalen mathematischen Kompetenzbereichen. Insbesondere für rechenschwache Kinder bietet das Goldmünzen-Scoring eine direkt ersichtliche Anreizstruktur, die schlechter Performanz aufgrund von Motivationsdefiziten vorbeugen kann, sowie die Entwicklung von zählenden hin zu abrufbasierten Rechenstrategien fördert. Aus diesen Gründen ist auch eine Implementation des Verfahrens in den inklusiven Unterricht denkbar.

Schlagwörter: Lernverlaufsdiagnostik, Mathematik, computergestützte Diagnostik, Speed-accuracy-tradeoff, Scoring

We are going on a gold coin hunt: Psychometric properties of different scorings in computer-based progress monitoring of mathematics ability

Abstract

Based on the robust indicator approach, a new progress monitoring instrument was developed and embedded into an online training for children with mathematical learning difficulties. The test captures the development in two basic mathematical competences: arithmetic fact knowledge (addition and subtraction up to 20) and numerical order processing (up to 100). According to the "high speed, high stakes" principle, speed and precision performance are combined into a single efficiency score on item level, expressed as the earnings or losses of gold coins. In a field study with primary school children (N = 241, grades 2 to 4), this new efficiency scoring showed both a higher reliability (r = .87-.93) and criterion validity (r = .51) than simple speed or precision scorings. Moreover, individual results in the progress monitoring test were sensitive to the sample's performance gains related to the training: For all three scorings, parameters of individual progress trajectories (random intercepts and random slopes) were predictive for post-training performance. Taken together, the new progress monitoring test qualifies as a reliable and valid tool for the formative assessment of primary school children's learning progress in basic mathematical abilities. Especially for low achieving children, the gold coin scoring offers an attractive incentive that should prevent low performance due to low motivation and foster the utilization of retrieval-based solution strategies. Hence, the test and training system could be implemented into remedial classroom practice.

Key words: learning progress monitoring, mathematics, computer-based diagnostics, speed-accuracy tradeoff, scoring

Spätestens seitdem der aus den USA stammende *Response-to-Intervention-Ansatz* (RTI) auch in Deutschland theoretisch diskutiert und modellhaft in die Schulpraxis implementiert wurde (Voß et al., 2016), besteht Einigkeit darüber, dass individualisierte Prävention und Förderung nur durch regelmäßige Verlaufsdiagnostik gesteuert werden kann. Historisch wurzelt das Konzept der Lernverlaufsdiagnostik in der Abgrenzung einer solchen formativen von einer summativen, statusdiagnotischen Evaluation. Diese Unterscheidung wird bereits seit einigen Jahrzehnten in der pädagogisch-psychologischen Forschung getroffen (Klauer, 2014). Während Statusdiagnostik Klassifikations- und Selektionsentscheidungen ermöglicht, dient die Lernverlaufsmessung der „Dokumentation des Lernfortschritts im Verlauf der Zeit" (Klauer, 2006, S. 17), sowie als Grundlage für die Planung von Fördermaßnahmen. Dabei ist es durchaus interessant, die Lernentwicklung gesamter, heterogener Schulklassen formativ hinsichtlich der durch die Lehrpläne definierten Kompetenzen zu messen. So kann die Entwicklung einzelner Schülerinnen und Schüler sowohl kriteriumsorientiert – an den Lernzielen – als auch im Hinblick auf eine soziale oder auch individuelle Bezugsnorm „evaluiert" werden. Hierfür existiert bereits eine überschaubare Menge publizierter deutschsprachiger Lernverlaufstests für den Bereich Lesen (computergestützt: Souvignier, Förster & Salaschek, 2014) und Rechnen (paper and pencil: Strathmann & Klauer, 2010, 2012; computergestützt: Souvignier et al., 2014). Bedenkt man die sonderpädagogischen Hintergründe von Lernverlaufsdiagnostik, die im englischsprachigen Bereich als *Curriculum-based Measurement* (CBM; Deno, 1985) bezeichnet wird, dann sollten „Breitbandtests" für gesamte Klassenstufen inhaltlich in zweierlei Hinsicht ergänzt werden: erstens bezüglich der Zielgruppe, an der die Verfahren ausgerichtet sind, und zweitens bezüglich der Kompetenzen, die für die Entwicklung ebendieser Zielgruppe

prädiktiv sind. Das bedeutet, dass Lernverlaufsdiagnostika frühe, valide und robuste Indikatoren für den sonderpädagogisch relevanten Leistungsbereich abdecken sollten, für die sich Aufgabenmengen anhand klarer Konstruktionsregeln definieren lassen. Eine bedeutende Zielgruppe mit Förderbedarf bilden dabei Grundschulkinder mit einer Entwicklungsstörung oder -schwäche im Lesen, Rechtschreiben oder in Mathematik.

Die Definition der relevanten Aufgabenmenge(n) ist auf zwei Wegen möglich (Fuchs, 2004). Ein Ansatz ist das *curriculum sampling*, wonach repräsentativ Aufgaben aus dem klassenspezifischen Curriculum abgeleitet werden. Dieses deduktive Vorgehen zeichnet sich durch eine hohe Lehrzielvalidität aus und ist in dieser Hinsicht für Lehrkräfte unmittelbar informativ. Demgegenüber steht der *robust indicator*-Ansatz, nach dem auch der hier vorgestellte Lernverlaufstest „Goldmünzenjagd" konstruiert wurde. Die Auswahl der Testaufgaben erfolgt dabei stärker induktiv als beim *curriculum sampling*. Es werden Kompetenzen bzw. Aufgaben ausgewählt, die sich empirisch als prädiktiv valide für die Gesamtleistung im interessierenden Bereich erwiesen haben. Der größte Vorteil des *robust indicator*-Ansatzes, der sich daraus ergibt, ist seine Flexibilität. Derart konstruierte Tests sind nahtlos über Klassenstufen hinweg einsetzbar und bieten deshalb das Potenzial einer Differenzierung im Anfangsunterricht, sowie fortlaufend in den unteren Leistungsbereichen (Foegen, Jiban & Deno, 2007; vgl. Walter, 2010 bzw. Walter, 2013 für Testverfahren der Lesegeschwindigkeit und des Leseverständnisses). Motivational spielt das Konzept des *Overlearnings* bei solchen eher einfachen Aufgaben eine wichtige Rolle. Es meint die Verbesserung der Sicherheit und Geschwindigkeit bei bereits ausgeprägter Präzision, um die Lernfreude lernschwacher Kinder zu fördern (Klauer, 2006). In Überblicksartikeln (Fuchs, 2004; Foegen et al., 2007) werden beide Konstruktionsweisen (*curriculum sampling* und *robust indicator approach*) unter dem Oberbegriff „CBM" zusammengefasst. Allerdings ist die Begriffswahl nicht ganz eindeutig, worauf auch in der Literatur verwiesen wird (Klauer, 2006; Klauer, 2014).

Der internetbasierte Lernverlaufstest „Goldmünzenjagd", der in diesem Beitrag vorgestellt wird, orientiert sich an basalen Kompetenzen, die für Kinder mit besonderem Förderbedarf in *Mathematik* eine Herausforderung darstellen. In der Literatur werden drei verbreitete Typen von CBM-Verfahren, die verschiedene Leistungsbereiche der Primarstufenmathematik erfassen, unterschieden: arithmetisches Basiswissen (z. B. Zählen, Mengenvergleich, fehlende Zahlen in Zahlenreihen ergänzen), die Grundrechenarten sowie Anwendungsaufgaben (Schätzen und Rechnen mit Größen, Sachaufgaben; Hosp, Hosp & Howell, 2007; Voß, 2016). Insbesondere Kompetenzen aus den ersten beiden Bereichen haben sich als frühe, robuste Indikatoren für die Entwicklung arithmetischer Leistung erwiesen. Aus dem Bereich des arithmetischen Basiswissens wird die ordinale Zahlenverarbeitung über die Klassenstufen 1-6 hinweg zum erklärungsstärksten basisnumerischen Prädiktor für die arithmetische Leistung (Lyons & Ansari, 2015; Lyons, Price, Vaessen, Blomert & Ansari, 2014). Diese Kompetenz, die sich im Grundschulalter entwickelt, ist eng mit dem Verstehen von Zahlensymbolen verknüpft. Sie wird wiederum bedingt durch den individuellen Entwicklungsgrad des Stellenwertverständnisses. Ein ausgereiftes Verständnis des Stellenwertprinzips und der sequentiellen Bündelung von Einern und Zehnern ist nicht zwingend nötig, um ordinale Vergleiche zweistelliger Zahlen vornehmen zu können. Solche Vergleiche sind aber besonders dann korrekt und effizient möglich, wenn ein sicheres Unterscheiden von Zehnern und Einern gelingt und gleichzeitig in eine symbolische Stellenwertnotation transkodiert werden kann (Fuson et al., 1997).

Im Bereich der Grundrechenarten ist das Speichern und der Abruf von einfachen Rechenfakten aus dem Langzeitgedächtnis be-

deutend für die Entwicklung effizienter nicht-zählender Rechenstrategien. Vor diesem Hintergrund ist die große interindividuelle Varianz im Faktenwissen von Grundschulkindern – sowohl in der Gesamtpopulation (Jordan, Hanich & Kaplan, 2003) als auch unter Kindern mit einer Rechenstörung (Geary, 2004; Geary, Hoard & Bailey, 2012) – und die Persistenz von schwachen, durchschnittlichen und effizienten Verarbeitungsprofilen im Faktenabruf (Vanbinst, Ceulemans, Ghesquière & De Smedt, 2015) beachtlich. Diese Fähigkeitsbereiche bilden wichtige Ansatzpunkte für evidenzbasierte Interventionsbausteine (vgl. Kaufmann, Handl & Thöny, 2003; Fuchs et al., 2009; Powell, Fuchs, Fuchs, Cirino & Fletcher, 2009; Wißmann, Heine, Handl & Jacobs, 2013 für Trainings des Faktenabrufs) – und damit kann Lernverlaufsdiagnostik, die solche grundlegenden *robust indicators* erfasst, zu einem Instrument für die Evaluation von Interventionseffekten jenseits des inhaltlich vorangeschrittenen Regelunterrichts werden.

Zusammenfassend haben Aufgaben aus den Gruppen arithmetisches Basiswissen und Grundrechenarten, neben den empirischen Belegen für ihre prädiktive Validität für die arithmetische Leistung, gegenüber der Gruppe der Anwendungsaufgaben (Hosp et al., 2007) den Vorteil der Auswertungsökonomie und -objektivität. Verglichen mit Sachaufgaben sind die Anforderungen eindimensionaler. Es wird also nicht – bzw. in geringerem Maße – zusätzlich Modellierungs- und Sprachkompetenz vorausgesetzt, die für das Lösen von Sachaufgaben notwendig ist. Dadurch sind sie auch für Kinder mit komorbiden Lernschwierigkeiten aussagekräftig.

Technisch wird die Qualität von Lernverlaufstests an den Anforderungen einer einfachen Wiederholbarkeit, identischen Schwierigkeit, hohen Durchführungs- und Auswertungsökonomie und psychometrischer Güte gemessen. Diese Kriterien sind vor allem dann gut umsetzbar, wenn das Generieren, Administrieren und Auswerten der Tests computergestützt erfolgt, und wenn mithilfe systematischer Konstruktionsregeln flexibel hypothetisch beliebig viele zufällige Aufgabenstichproben – und damit schwierigkeitshomogene Tests – erzeugt werden können. Testtheoretische und psychometrische Probleme, die scheinbar durch ein zufälliges *item sampling* auftreten, d. h. dadurch, dass jedes Individuum zu Testzeitpunkt *t* eine eigene Itemstichprobe bearbeitet, lassen sich durch eine generische Testkonzeption auflösen, nach der Lösungskompetenz nicht mehr auf Grundlage von einzelnen Aufgaben, sondern auf Grundlage von Aufgaben*typen* definiert wird (Rohwer, 2015). Im Falle einer vollständig zufälligen Itemauswahl auf individueller Ebene können derartig konstruierte Tests nach dem Binomialmodell ausgewertet werden (Klauer, 2011), bei dem die Personenfähigkeit dem Anteil korrekt gelöster Aufgaben entspricht.

Je nach Testkonstruktion bieten sich verschiedene Scorings an, um die Leistung in Lernverlaufstests zu bewerten. Wenn es sich um Power-Tests handelt (vgl. LVD-M 2-4, Strathmann & Klauer, 2012), ist vor allem die Präzision von Interesse, die standardmäßig als Anteil korrekter Antworten an der Menge der bearbeiteten Aufgaben definiert wird. Bei Speed-Tests lässt sich zusätzlich die Anzahl bearbeiteter Aufgaben, also die Geschwindigkeit, interpretieren. Beide Scoring-Varianten betonen *einen* Leistungsaspekt. Effizienz-Maße hingegen kombinieren die Aspekte. Sie können als Präzisionsmaß, das für die Bearbeitungsgeschwindigkeit gewichtet wird, aufgefasst werden und haben somit den Vorteil der Sparsamkeit, da sie beide Informationen zu einem Kennwert verdichten. Damit geht als Limitation einher, dass keine differenzierten Aussagen zu individuellen *speed-accuracy tradeoffs* getroffen werden können, sodass die klassischen Maße zusätzliche diagnostische Informationen beitragen können. Ein solcher Effizienz-Index, die Bearbeitungsflüssigkeit, wurde von Voß (2016) für ein CBM-Instrument mit Additions- und Subtraktionsaufga-

ben im Zahlenraum bis 20 nach einer Formel bestimmt, die die Präzision im Vergleich zur Geschwindigkeit im doppelten Maße gewichtet. Der von Voß (2016) berechnete Index bezieht sich auf die gesamte Aufgabenmenge. Eine Möglichkeit, die Bearbeitungseffizienz bereits auf Itemebene zu bewerten, ergibt sich aus dem sogenannten *high speed, high stakes*-Scoring (HSHS; Klinkenberg, Straatemeier & van der Maas, 2011; Maris & Van der Maas, 2012). Testtheoretisch sind aus dieser Perspektive schnelle Antworten informativer als langsam gegebene Antworten (Maris & Van der Maas, 2012). Im HSHS-Ansatz müssen Versuchspersonen die Aufgaben nicht nur korrekt, sondern gleichzeitig auch schnell und somit effizient bearbeiten, damit eine hohe Fähigkeitsausprägung angenommen wird. Deshalb werden schnelle Richtigantworten durch mehr Punkte belohnt als langsame, und schnelle Falschantworten (z. B. durch Raten) durch höheren Punktabzug bestraft als langsamere. Im hier vorgestellten Lernverlaufstest „Goldmünzenjagd" erfolgt diese Verrechnung von Geschwindigkeit und Präzision während der Testbearbeitung und ist für die teilnehmenden Kinder durch den Gewinn oder Verlust von Münzen sichtbar. Ein vergleichbares Scoring wurde bereits in dem niederländischen adaptiven Test- und Trainingssystem *The Maths Garden* umgesetzt (Klinkenberg et al., 2011).

Zusammenfassend werden mit diesem Beitrag zwei Ziele verfolgt: erstens die Vorstellung eines computergestützten *robust-indicator*-basierten Lernverlaufstests, der sich für die interventionsbegleitende Diagnostik rechenschwacher Kinder eignet. Dabei soll die HSHS-Scoring-Methode, die Geschwindigkeit und Präzision kombiniert, beschrieben und psychometrisch mit klassischen Scorings verglichen werden. Der Lernverlaufstest ist in ein computergestütztes Training für rechenschwache Kinder eingebettet. Deshalb soll zweitens die Prädiktivität der damit gemessenen Lernverläufe für die Leistungsentwicklung der Grundschulkinder, die an dem Training teilgenommen haben, untersucht werden. Diese Analysen sind verwandt mit dem Konzept der Änderungssensitivität (Klauer & Strathmann, 2013). Hierfür wird neben der Gesamtgruppe auch eine Teilgruppe der Kinder betrachtet, deren mathematische Leistung im basisnumerischen und arithmetischen Eingangsscreening im unteren Normquartil[1] und damit im Risikobereich für eine Rechenschwäche lag.

Methode

Der computergestützte Lernverlaufstest „Goldmünzenjagd" ist an ein Online-Training für Kinder mit Rechenschwierigkeiten gekoppelt (Kuhn & Holling, 2014). Als Kriterium für die Validierung diente ein computergestütztes, statusdiagnostisches Screening (CODY; Kuhn, Raddatz, Holling & Dobel, 2013). Dieser Test wurde zu Beginn des Trainings durchgeführt ($CODY_{prä}$) und nach Abschluss der selbst gewählten Trainingsdauer (M = 30.94, SD = 0.94 Trainingstage) wiederholt ($CODY_{post}$).

Online-Training

Das Online-Training (Kuhn & Holling, 2014), das konzeptuell an das Screening anknüpft, kombiniert Aufgaben aus sieben teils überlappenden Bereichen: Zahl-Größen-Verknüpfung, Zahlenstrahlschätzaufgaben, Teil-Ganzes-Verständnis, Dezimalsystem/Transkodieren, mathematisches Faktenwissen und Rechnen, Mathematisieren/

[1] Das untere Normquartil, d. h. PR ≤ 25, liegt oberhalb des Cut-Off-Wertes, der im klinischen Kontext zur Identifikation von Kindern mit einer Lernstörung angelegt wird. Dort wird i. d. R. eine Schwelle von -1 SD (PR ≤ 16), -1.5 SD (PR ≤ 7) oder noch geringer gewählt. Im Forschungskontext ist die PR-25-Schwelle durchaus gängig. Neben der klinisch relevanten Gruppe wird dadurch eine Risikogruppe („low achievers", meist 11 ≤ PR < 25, z. B. Geary, 2013) miteingeschlossen. Dies erhöht die Stichprobengröße und damit auch die Power, d. h. statistische Belastbarkeit, von Analysen.

Textaufgaben sowie Arbeitsgedächtnis. Die Kinder bearbeiten pro Trainingseinheit zwei dieser Aufgaben jeweils 10 Minuten lang und verbringen zusätzliche 10 Minuten mit einer Rahmengeschichte in der Phantasiewelt Talasia. Das Training ist adaptiv, sodass jedem Kind eingangs auf Grundlage seines Ergebnisses im CODY-Screening eines von vier spezifischen Testprofilen zugeordnet wird: basisnumerisch, Rechnen, Arbeitsgedächtnis oder ausgeglichen. Daraus ergibt sich eine schwerpunktmäßige Auswahl von Aufgaben, die die im Screening ermittelten Defizite trainieren. Innerhalb der einzelnen Aufgaben wird das Schwierigkeitslevel blockweise adaptiert, sodass jedes Kind Items der Schwierigkeit erhält, für die es eine mittlere Lösungsquote von etwa 80%, welche für den Lernerfolg günstig ist (Jansen et al., 2013), erreicht. Die Anmeldung erfolgt über eine Online-Plattform (www.meistercody.com/de/talasia) und ist kostenpflichtig. Standardmäßig wird ein Trainingsumfang von 30 Einheiten (fünfmal pro Woche über einen Zeitraum von 6 Wochen hinweg) empfohlen, danach kann jedoch beliebig weiter trainiert werden. Wirksamkeitsnachweise für frühere (Kuhn & Holling, 2014) sowie die aktuelle Version des Trainings (Kuhn, 2016; Kuhn et al., 2017) liegen in Form von Evaluationsstudien mit Kontrollgruppen und basisnumerischen sowie curricularen und nicht-curricularen Kriterien vor.

Statusdiagnostik

Das CODY-Screening, das in dieser Studie zur Statusdiagnostik eingesetzt wurde, deckt vier faktorenanalytisch bestätigte Leistungsbereiche ab, die zu einem Gesamtscore integriert werden: basale Zahlenverarbeitung (Mengenvergleiche, Zählen), komplexe Zahlenverarbeitung (Zahlenstrahl, Zahlensteine, Transkodieren, fehlende Zahl in Zahlenreihen), Rechnen (Addition, Subtraktion, Multiplikation, Platzhalteraufgaben) und Arbeitsgedächtnis (Matrixspanne). Die Retest-Reliabilität nach zwei Wochen beträgt r_{tt} = .88. Es liegen klassen- und halbjahresspezifische Normen für den Gesamt- und die vier Skalenwerte vor. Für eine nähere Beschreibung der Testaufgaben wird auf die Studien von Kuhn et al. (2013) bzw. Raddatz, Kuhn, Holling, Moll und Dobel (2016) verwiesen. Den Erziehungsberechtigten der teilnehmenden Kinder wurde empfohlen, das Screening als Einstufungstest vor Beginn und als Abschlusstest am Ende der Trainingszeit durchzuführen. Die Testzeitpunkte waren jedoch frei wählbar. Dies führt dazu, dass den einzelnen Analysen, die in diesem Artikel berichtetet werden, unterschiedliche Stichproben zugrunde gelegt wurden (s. Abschnitt Stichprobe).

Lernverlaufstest

Der Lernverlaufstest, der das erste Mal vor der fünften Einheit des oben beschriebenen Trainings automatisch startet, folgt der Story einer Goldmünzenjagd. Die hier verdienten Goldmünzen können die Kinder am Ende der Trainingseinheit als Zahlungsmittel zur Gestaltung ihres virtuellen Palastgartens einsetzen. Die Verlaufstests erfolgen im Abstand von fünf Trainingssessions, was einer Frequenz von einer Testung pro Woche entspricht (vgl. Fuchs, Fuchs, Hamlett, Phillips & Bentz, 1994; Fuchs, Compton, Fuchs, Paulsen, Bryant & Hamlett, 2005), wenn wie empfohlen trainiert wird. Sie setzen sich stets aus drei repräsentativen Aufgabentypen zusammen, die unabhängig von klassenstufenspezifischen Curricula als *robust indicators* der arithmetischen Entwicklung belegt sind: Additionsfakten, Subtraktionsfakten und Zahlenordnen. Alle drei Aufgabentypen werden nacheinander zu Beginn des Tests anhand von Beispielen erläutert, dann folgen ohne Unterbrechung die Trials, die jeweils durch einen Fixationsstern (500 ms) getrennt werden. Die reine Testzeit pro Aufgabentyp beträgt 90 Sekunden, insgesamt also viereinhalb Minuten. Die *Additions- und Subtraktionsaufgaben* bestehen aus zwei Teilmengen: a) ohne Zehnerübergang (E + E = E und E - E = E bzw. E + E

= 10 und 10 - E = E) und b) mit Zehnerübergang im Zahlenraum bis 20 (E + E = ZE und ZE - E = E), die durchmischt präsentiert werden. Die Antworteingabe erfolgt per Tastatur oder Anklicken einer Ziffernleiste, eine Korrektur einmal eingegebener Werte ist nicht möglich. Die Items des Subtests *Zahlenordnen* bestehen aus jeweils einem zweistelligen Zahlentriple, für das angegeben werden soll, ob es korrekt geordnet ist oder nicht (Lyons & Ansari, 2015). Es kommen drei unterschiedliche Itemtypen vor: kongruente, für die der Zehner immer größer ist als der Einer (z. B. 32, 61, 54), inkongruente, die ein bis zwei Elemente mit Z < E enthalten (z. B. 23, 71, 45), und Kontrollitems, für die alle drei Zehner identisch sind (z. B. 42, 46, 49). Die Aufgabentypen werden randomisiert präsentiert, Kontrollitems jedoch seltener (jedes dritte bis fünfte Item). Die Hälfte der Triple im Aufgabenpool ist richtig, die andere Hälfte falsch geordnet. Die Eingabe erfolgt, indem ein Häkchen oder ein Kreuz angeklickt wird.

Im Sinne einer generischen Testkonstruktion (Rohwer, 2015) werden die Items pro Testzeitpunkt und Kind zufällig aus dem durch die Konstruktionsregeln definierten Pool möglicher Aufgaben gezogen. Bei Additions- und Subtraktionsaufgaben ist es aufgrund der beschränkten Aufgabenmenge folglich hypothetisch möglich, dass sich einzelne Items wiederholen, allerdings erst dann, wenn innerhalb der beschränkten Testzeit alle konstruierten Aufgaben bereits einmal vorgegeben wurden. Im Anschluss an jeden Lerntest erhalten die Erziehungsberechtigten eine Ergebniszusammenfassung per Mail (s. Abbildung 1).

Scorings

Klassischerweise wird für jedes Kind pro Aufgabentyp und Testtag die Anzahl bearbeiteter Aufgaben und der Anteil korrekter Lösungen bestimmt. In dieser Studie wird zusätzlich die Effizienz durch ein HSHS-Scoring bewertet, das die Kinder live für jede bearbeitete Aufgabe durch einen Gewinn oder Verlust von Goldmünzen nachvollziehen können (s. Abbildung 1). Durch dieses Münzscoring werden Antworten mit hohem Informationswert, d. h. schnelle Richtigantworten (z. B. effizienter Faktenabruf) oder schnelle Falschantworten (z. B. Raten) besonders stark gewichtet. Richtigantworten werden direkt mit dem Gewinn von Münzen belohnt, Falschantworten sind mit einem entsprechenden Münzverlust verbunden, sodass schnelles Raten negative Konsequenzen hat. Die Staffelung der Münzgewinne bzw. -verluste abhängig von der Antwortzeit beginnt mit fünf Münzen bei einer Zeitgrenze von 3.5 Sekunden, bis zu der von Faktenabruf aus dem Langzeitgedächtnis ausgegangen wird (Andersson, 2010), und wird in 3-Sekunden-Intervallen fortgesetzt. Für eine richtige Antwort in höchstens 8 Sekunden würden beispielsweise drei Münzen gutgeschrieben, für eine Falschantwort nach 5 Sekunden vier Münzen abgezogen. Antwortzeiten von 12.5 Sekunden oder länger entsprechen einer Münze. Als Score wird die Summe aus Münzgewinnen und -verlusten über alle bearbeiteten Items hinweg bestimmt, wobei sehr selten auftretende Negativscores in der internen Rechnung berücksichtigt werden, im Endergebnis für die Kinder jedoch auf Null gesetzt werden.

Stichprobe

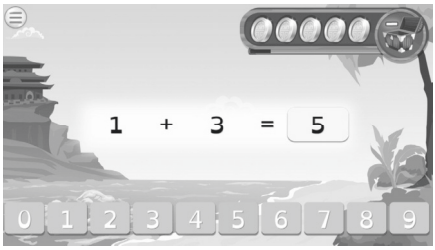
Es liegen Lernverlaufsdaten von insgesamt 241 Kindern der Klassenstufen 2 bis 4 vor (s. Tabelle 1), die im Zeitraum zwischen Oktober 2015 und August 2016 von ihren Erziehungsberechtigten für das CODY-Training angemeldet wurden. 233 Kinder (97%) absolvierten den Lernverlaufstest bis zum Ende der regulär vorgesehenen Trainingsdauer, d. h. bis Trainingstag 30. Danach fuhr nur ein kleiner Teil der Stichprobe mit der Intervention, und damit auch der Lernverlaufsdiagnostik, fort (n = 29[2] bis Trainingstag 35, n = 4 bis Trainingstag 50). Die

[2] Teilstichproben werden in diesem Beitrag mit n gekennzeichnet, die gesamte Stichprobe mit N.

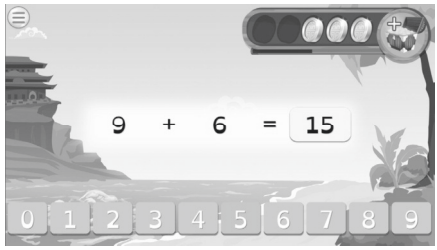
Reliabilität des Lernverlaufstests wurde mit der gesamten Stichprobe überprüft. Für die anschließenden Fragestellungen wurde die Stichprobe nach inhaltlichen Gesichtspunkten eingeschränkt. Um die Kriteriumsvalidität mit dem Statusdiagnostikum (CODY-Screening, s.o.) zu untersuchen, wurden nur die $n = 174$ Kinder ausgewählt, die das Einstufungsscreening ($CODY_{prä}$) innerhalb der ersten fünf Trainingseinheiten und damit in zeitlicher Nähe zum ersten Lernverlaufstest durchgeführt hatten.

Addition

a) ohne Zehnerübergang

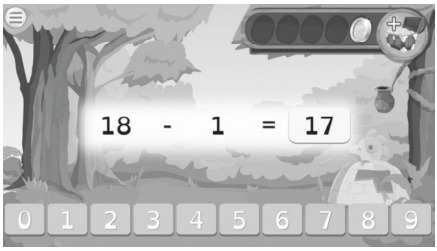


b) mit Zehnerübergang

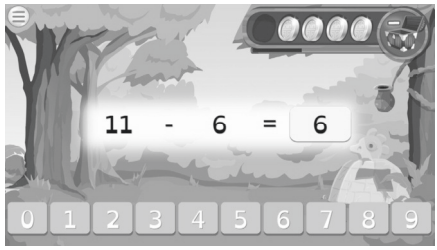


Subtraktion

a) ohne Zehnerübergang



b) mit Zehnerübergang



Zahlenordnen

a) kongruentes Item

b) inkongruentes Item

Feedback

a) spielintern

b) per E-Mail an die Erziehungsberechtigten

[…] Für jede Antwort kann Musterkind maximal fünf Münzen gewinnen. Je langsamer dein Kind antwortet, desto weniger Münzen kommen hinzu. Für eine falsche Antwort werden sogar Münzen abgezogen. […]

*Heute hat Musterkind **xx** Aufgaben bearbeitet. Davon wurden **xx** richtig beantwortet. Dies entspricht einem Anteil richtiger Antworten von **xx** Prozent. Dafür gab es **xx** Münzen. Durchschnittlich beantworten Grundschulkinder etwa 23 Aufgaben beim Lerntest richtig und erhalten dafür 58 Münzen.*

Abbildung 1: Screenshots zum Goldmünzen-Scoring für alle drei Aufgabentypen und Feedback

Tabelle 1: Stichprobe

	Gesamt-Stichprobe (N = 241)	Zeitplankonforme Stichprobe[a]	
		Gesamt (n = 127)	PR ≤ 25 (n = 33)
Geschlecht (% Mädchen)	63.90%	71.65%	78.79%
Klassenstufe (2/3/4)	(152/58/31)	(86/24/17)	(18/10/5)

Anmerkungen: [a] Einstufungstest ($CODY_{prä}$) in den ersten 5 Trainingseinheiten, Abschlusstest ($CODY_{post}$) 25-30 Einheiten nach Prätest; PR ≤ 25: Subgruppe mit einem Prozentrang ≤ 25 im CODY-Prätest.

Für die Analysen zur Änderungssensitivität und zur Prädiktivität der Lernverläufe für die Leistungsentwicklung wurden alle n = 127 Kinder berücksichtigt, die das Einstufungs- *und* Abschlussscreening zeitplankonform durchlaufen hatten (im Folgenden: zeitplankonforme Stichprobe). Ausgewählt wurden hierfür wie bei den Analysen zur Kriteriumsvalidität diejenigen Kinder, deren Einstufungstest in den ersten fünf Trainingstagen erfolgte. Zusätzlich wurden nur die Fälle eingeschlossen, in denen der Abschlusstest ($CODY_{post}$) – wegen der empfohlenen Standarddauer von 30 Einheiten – 25 bis 30 Trainingstage später erfolgt war. Diese Kinder bearbeiteten im ersten Lerntest im Mittel 29.04 (SD = 7.36) Aufgaben, davon 73.83% (SD = 14.69%) korrekt und verdienten dafür 58.72 (SD = 46.00) Münzen. Schließlich wurden diese Analysen mit einer leistungsschwachen Teilstichprobe (n = 33), die beim Einstufungstest ein Ergebnis im unteren Normquartil (PR ≤ 25) erzielt hatte, wiederholt. In dieser Subgruppe wurden im ersten Lerntest im Mittel 25.06 (SD = 7.45) Aufgaben bearbeitet, davon 64.44% (SD = 15.54%) korrekt, für insgesamt 28.58 (SD = 25.82) Münzen.

Statistische Auswertung

Für alle hier beschriebenen Analysen wurde die Statistiksoftware *R* (Version 3.3.2; R Core Team, 2016) mit den unten genannten Paketen verwendet. In einem *ersten* Schritt wurden die Lernverläufe pro Scoring grafisch dargestellt. Innerhalb der Scorings wurde zwischen den drei Subtests (Addition, Subtraktion, Zahlenordnen) differenziert, um zu untersuchen, inwiefern qualitative Unterschiede zwischen den Anforderungsbereichen bestehen. In einem *zweiten* Schritt wurden die drei verschiedenen Scorings (Geschwindigkeit, Präzision und Effizienz) psychometrisch, bezüglich ihrer Reliabilität und Kriteriumsvalidität verglichen. Dabei wird zwischen Subtests, und für die Kriteriumsvalidität auch zwischen Klassenstufen, differenziert. Für den Anteil korrekter Antworten (Präzision) und die Anzahl gewonnener Münzen (Effizienz) wurde die Spearman-Brown-korrigierte Split-Half-Reliabilität mit Odd-Even-Split berechnet. Da für die Anzahl bearbeiteter Items (Geschwindigkeit) keine sinnvolle Split-Half-Reliabilität bestimmt werden kann, wird hierfür die Retest-Reliabilität für die jeweils aufeinander folgenden Testzeitpunkte, also z. B. für Trainingstag 5 und 10, berichtet. Die Kriteriumsvalidität wurde als die Korrelation der Scores im ersten Lernverlaufstest mit dem statusdiagnostischen Einstufungsscreening ($CODY_{prä}$) bestimmt. In einem *dritten* Schritt sollte die Änderungssensitivität des Lernverlaufstests untersucht werden. Dazu wurden zunächst für alle drei Scorings mit der Funktion *lmer* aus dem R-Paket *lme4* (Bates, Maechler, Bolker & Walker, 2015) unter Verwendung eines Restricted-Maximum-Likelihood-Schätzers (REML) separate Random-Intercept-Random-Slope-Modelle angepasst. Neben dem Gesamtscore wurden auch die Scores der Subtests als abhängige Variablen verwendet. Für jedes Kind resultierten Schätzungen des Intercepts (ri, individuelles Ausgangsniveau) und des Steigungsparameters (rs, individueller Lernverlauf) als bedingte Mittelwerte der

Tabelle 2: Variabilität (SD) und Korrelation der random effects

	Geschwindigkeit		Präzision		Effizienz	
	SD	***r*(ri, rs)**	***SD***	***r*(ri, rs)**	***SD***	***r*(ri, rs)**
ri_{gesamt}	6.85	-0.36**	0.12	-0.05	43.67	0.23
rs_{gesamt}	0.15		0.002		0.95	
Residuum	3.28		0.10		24.16	
ri_{Add}	2.64	-0.45**	0.16	-0.46	14.20	0.09
rs_{Add}	0.06		0.005		0.50	
Residuum	1.51		0.15		11.49	
ri_{Sub}	2.65	-0.30	0.18	-0.59	16.12	0.10
rs_{Sub}	0.04		0.004		0.36	
Residuum	1.78		0.19		13.36	
ri_{ZO}	2.06	-0.29	0.15	-0.31	19.49	-0.17
rs_{ZO}	0.06		0.003		0.45	
Residuum	1.60		0.12		13.91	

Anmerkung: ri = random intercept im Lerntest, rs = random slope; Add = Addition, Sub = Subtraktion, ZO = Zahlenordnen; *$p < .05$.,**$p < .01$

zufälligen Effekte. Die Standardabweichungen und Korrelationen der verschiedenen *random effects* sind in Tabelle 2 dargestellt.

Diese Parameter wurden schließlich, zusätzlich zur Leistung im Einstufungsscreening ($CODY_{prä}$), in vollstandardisierten multiplen Regressionsmodellen als Prädiktoren verwendet, um die Leistung im Abschlusstest ($CODY_{post}$) vorherzusagen. Dafür wurden schrittweise drei Modelle spezifiziert und hinsichtlich der inkrementellen Varianzaufklärung verglichen: Das Baseline-Modell 0, das nur die Leistung im Einstufungsscreening als Prädiktor enthält, Modell 1 mit den aggregierten, auf dem jeweiligen Gesamtscore des Lernverlaufstests basierenden *random intercepts* und *random slopes* sowie Modell 2, das diese Parameter für die Lernverlaufssubtests (Addition, Subtraktion, Zahlenordnen) differenziert:

Modell 0: $CODY_{post} \sim CODY_{prä}$

Modell 1: $CODY_{post} \sim CODY_{prä} + ri_{gesamt} + rs_{gesamt}$

Modell 2: $CODY_{post} \sim CODY_{prä} + ri_{Add} + rs_{Add} + ri_{Sub} + rs_{Sub} + ri_{ZO} + rs_{ZO}$

Um den relativen Beitrag der einzelnen Prädiktoren zur gesamten Varianzaufklärung zu schätzen, wurden anschließend pro Scoring Dominanzanalysen mit dem R-Paket *relaimpo* (Grömping, 2006) durchgeführt. Dabei werden alle möglichen Konstellationen der Prädiktoren in einer multiplen Regression berücksichtigt. In einem *vierten* explorativen Schritt wurden die Lernverläufe (*random slopes*) von zwei Teilgruppen derjenigen Kinder, die beim Einstufungsscreening im unteren Normquartil lagen, verglichen: eine Gruppe mit statusdiagnostischer Verbesserung und eine Gruppe ohne Verbesserung infolge des Trainings. Effekte einer Intervention (Verbesserung, Verschlechterung, keine Veränderung) können mithilfe des *Reliable Change Index* (RC; Jacobson & Truax, 1991; Jabrayilov, Emons & Sijtsma, 2016) klassifiziert werden: der Differenz aus Prä- und Posttest-Score, die an der Reliabilität des Tests standardisiert wird. Veränderungen, für die $|RC| \geq 1.645$ gilt (was einem zweiseitigen 10%-Signifikanzniveau entspricht), werden als statistisch bedeutsam angesehen (Jabrayilov et al., 2016). Ei-

ne klinische Verbesserung liegt vor, wenn das Ergebnis des Prätests in einen kritischen Bereich fällt (d. h. hier PR ≤ 25), das des Posttests hingegen aber nicht mehr, eine Verschlechterung im umgekehrten Fall. Statistische und klinische Veränderungen können, müssen aber nicht, gleichzeitig auftreten. Zum Beispiel wäre eine klinische Verbesserung auf ein Leistungsniveau außerhalb des definierten Risikobereichs möglich, ohne statistisch bedeutsam zu sein.

Ergebnisse

Lernverläufe

Bevor die Ergebnisse der Analysen zu Reliabilität, Validität und Änderungssensitivität der Scorings berichtet werden, sollen zunächst die Lernverläufe für die drei Scorings, und innerhalb jedes Scorings für die drei Subtests, betrachtet werden (Abbildung 2). Dabei wird die Gesamtgruppe und zu-

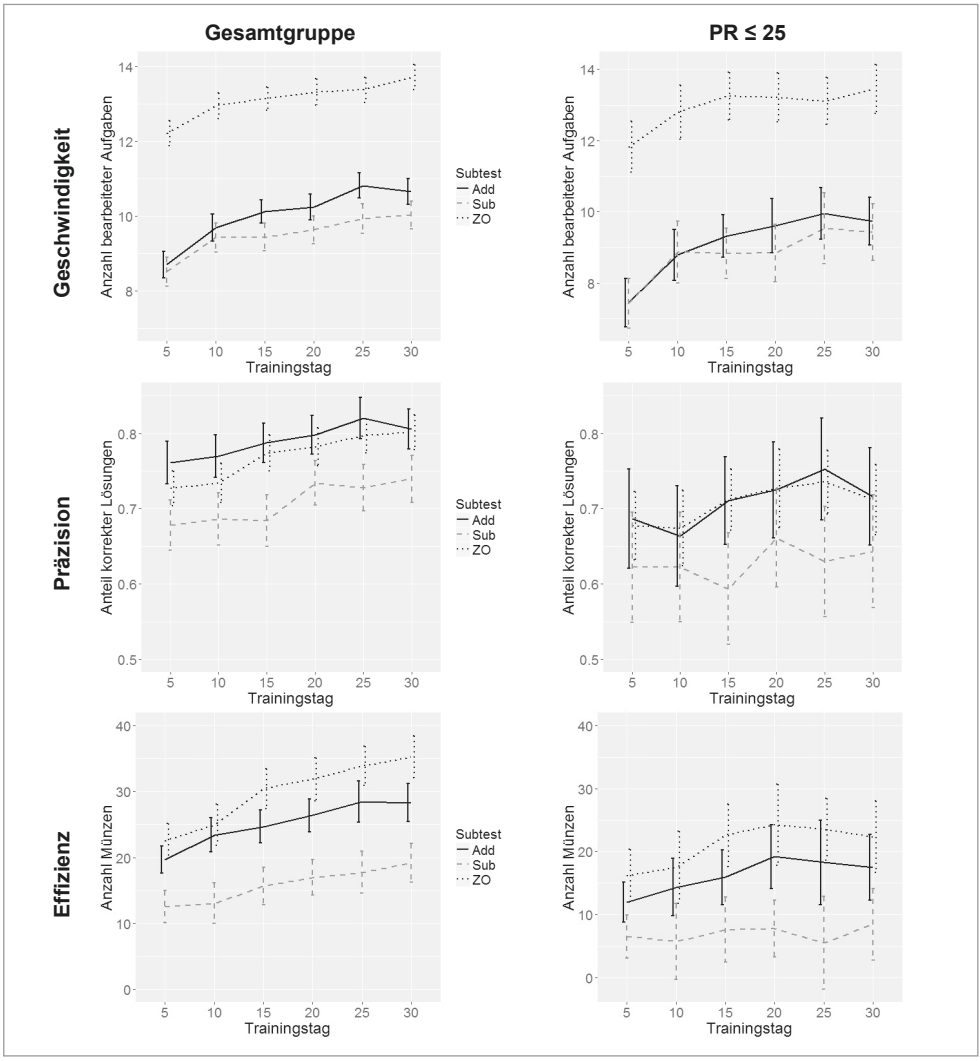


Abbildung 2: Geschwindigkeit, Präzision und Effizienz pro Subtest (Add = Addition, Sub = Subtraktion, ZO = Zahlenordnen) für die gesamte Stichprobe (N = 241) und die Subgruppe im unteren Leistungsquartil des Einstufungsscreenings ($CODY_{prä}$ PR ≤ 25, n = 65), Mittelwerte und 95%-Konfidenzintervalle pro Testzeitpunkt

sätzlich die leistungsschwächste Subgruppe (PR ≤ 25), d. h. die Zielgruppe des Tests und Trainings, dargestellt. Unter dem Geschwindigkeitsscoring zeigt sich eine Überlegenheit des Subtests Zahlenordnen. Sowohl die Gesamtgruppe als auch die Teilgruppe bearbeiteten mehr Aufgaben aus diesem Anforderungsbereich als aus den beiden Subtests zum arithmetischen Faktenwissen. Die Lernverläufe unterscheiden sich im Spezifischen zwar statistisch bedeutsam zwischen den Subtests ($p < .001$ in der Gesamtgruppe, aber $p > .05$ in der Subgruppe) [3]. Qualitativ betrachtet ähneln die Muster sich allerdings stark. Gleiches gilt auch für das Präzisionsscoring ($p < .001$ in der Gesamtgruppe, $p < .05$ in der Subgruppe) und Effizienzscoring ($p < .001$ in der Gesamtgruppe, aber $p > .05$ in der Subgruppe). Unter diesen beiden Scorings fallen die vergleichsweise geringen Werte im Bereich der Subtraktionsaufgaben auf. Sowohl die Gesamtgruppe als auch die Teilgruppe bearbeitete die Subtraktionsaufgaben im Vergleich zu den anderen beiden Subtests weniger präzise und weniger effizient. Für die leistungsschwache Subgruppe zeigen sich im Präzisions- und Effizienzscoring – anders als für die Geschwindigkeit – über alle Inhaltsbereiche hinweg zudem weniger konstante Lernzuwächse.

[3] Dies wurde mit einem Modellvergleich mittels Likelihood-Ratio-Test überprüft. Verglichen wurden je zwei Modelle, mit dem Ergebnis in der jeweiligen Scoring-Methode als abhängige Variable. Das Baseline-Modell enthält den Haupteffekt des Subtests (d. h. die konstante Über- oder Unterlegenheit in den einzelnen Inhaltsbereichen) gleichermaßen als festen und zufälligen Effekt. Es wurde verglichen mit einem Modell, das zusätzlich die Interaktion zwischen Subtest und Trainingstag (d. h. die Unterschiede in den Lernverläufen zwischen den Subtests) beinhaltet. Fällt der Test signifikant aus, dann kann von einem inkrementellen Effekt der Interaktion ausgegangen werden. Dieser entspricht einer Abhängigkeit des Lernverlaufs vom Subtest.

Psychometrische Analysen

Im Vergleich der drei Scorings ist das Münzscoring (Effizienz) am reliabelsten ($rel_{splithalf}$ = .87-.93, Tabelle 3) und weist die höchste Kriteriumsvalidität auf (r = .51). Im Vergleich der Subtests schneidet das Zahlenordnen in der Kriteriumsvalidität und in der Reliabilität unter dem Effizienzscoring vergleichsweise am schwächsten ab. Die Kriteriumsvaliditäten fallen innerhalb der einzelnen Jahrgangsstufen, insbesondere für Klasse 4, deutlich höher aus (.38-.85) als über die gesamte Stichprobe hinweg.

Änderungssensitivität

Im Folgenden werden pro Scoring die drei oben beschriebenen Modelle verglichen (Tabelle 4): Die Modelle mit den Lerntest-Variablen (*random intercepts* und *random slopes*) klären über das Baseline-Modell, das nur die Leistung im Einstufungsscreening ($CODY_{prä}$) enthält, hinaus zwischen 1.3% und 4.3% zusätzliche Varianz auf. Diese inkrementelle Varianzaufklärung fällt, außer für Modell 2 das Präzisionsscorings, signifikant aus. Die Dominanzanalyse (Grömping, 2006) zeigt, dass für alle drei Scorings die Leistung im Einstufungsscreening der stärkste Prädiktor (Geschwindigkeit: 36.19%, Präzision: 28.52%, Effizienz: 25%) ist. Unter dem Präzisionsscoring ist die Eingangsleistung im Lerntest (*random intercept*) der zweitwichtigste Prädiktor (14.89%), und die Entwicklung im Lerntest (*random slope*) der drittwichtigste (2.44%). Unter den Geschwindigkeits- und Effizienzscorings unterscheiden sich *random intercept* (7.20% bzw. 12.29%) und *random slope* (3.72% bzw. 10.10 %) nicht statistisch signifikant in ihrem Beitrag zur Varianzaufklärung. Unter den einzelnen Lernverlaufssubtests (Addition, Subtraktion, Zahlenordnen) zeigten sich keine herausstechenden Prädiktoren, sodass die Modellparameter der Modelle 2 nicht berichtet werden.

In einem letzten Schritt wurde explorativ für die leistungsschwache Subgruppe

Tabelle 3: Psychometrischer Vergleich der Scorings

	Reliabilität (N = 241)						Kriteriumsvalidität (n = 174)
Lerntest Nr.	**1**	**2**	**3**	**4**	**5**	**6**	**Korrelation Lerntest 1 mit $CODY_{prä}$**
Geschwindigkeit (gesamt)	-	*.79*	*.78*	*.76*	*.76*	*.80*	*.43 (K_2: .51, K_3: .50, K_4: .79)*
Add	-	.66	.66	.64	.68	.72	.44 (K_2: .52, K_3: .38, K_4: .70)
Sub	-	.67	.69	.62	.66	.69	.39 (K_2: .51, K_3: .50, K_4: .79)
ZO	-	.66	.62	.65	.64	.65	.28 (K_2: .33, K_3: .33, K_4: .59)
Präzision (gesamt)	*.72*	*.84*	*.82*	*.78*	*.86*	*.83*	*.39 (K_2: .45, K_3: .38, K_4: .60)*
Add	.50	.64	.69	.62	.69	.73	.32 (K_2: .38, K_3: .26, K_4: .46)
Sub	.49	.70	.66	.66	.66	.68	.28 (K_2: .30, K_3: .34, K_4: .34)
ZO	.56	.68	.61	.71	.66	.68	.22 (K_2: .25, K_3: .11, K_4: .42)
Effizienz (gesamt)	*.87*	*.91*	*.90*	*.90*	*.93*	*.90*	*.51 (K_2: .61, K_3: .58, K_4: .85)*
Add	.82	.86	.86	.83	.88	.89	.47 (K_2: .54, K_3: .43, K_4: .77)
Sub	.81	.87	.84	.82	.90	.83	.41 (K_2: .45, K_3: .42, K_4: .68)
ZO	.69	.73	.71	.79	.75	.76	.31 (K_2: .33, K_3: .35, K_4: .57)

Anmerkung: Die Kriteriumsvalidität wurde mit der Teilgruppe, die das Screening in den ersten 5 Trainingseinheiten und damit in zeitlicher Nähe zum ersten Lernverlaufstest absolviert hatte, bestimmt; Add = Addition, Sub = Subtraktion, ZO = Zahlenordnen, K_2 = Klasse 2 (n = 114), K_3 = Klasse 3 (n = 38), K_4 = Klasse 4 (n = 22)

Tabelle 4: Regressionsmodelle zur Vorhersage der Leistung nach dem Training (n = 127)

	β	SE	t
Baseline			
Modell 0 ($R^2_{adjustiert}$ = 0.418)			
$CODY_{prä}$	0.65***	0.07	9.57
Anzahl bearbeiteter Aufgaben (Geschwindigkeit)			
Modell 1 ($R^2_{adjustiert}$ = 0.458; ΔR^2 = 0.040**)			
$CODY_{prä}$	0.63***	0.08	8.23
ri_{gesamt}	0.10	0.08	1.22
rs_{gesamt}	0.22**	0.07	3.31
Modell 2 ($R^2_{adjustiert}$ = 0.450; ΔR^2 = 0.032**)			
Anteil korrekter Lösungen (Präzision)			
Modell 1 ($R^2_{adjustiert}$ = 0.445; ΔR^2 = 0.027*)			
$CODY_{prä}$	0.51***	0.08	6.16
ri_{gesamt}	0.22*	0.09	2.48
rs_{gesamt}	0.04	0.07	0.48
Modell 2 ($R^2_{adjustiert}$ = 0.431; ΔR^2 = 0.013)			
Anzahl Münzen (Effizienz)			
Modell 1 ($R^2_{adjustiert}$ = 0.461; ΔR^2 = 0.043***)			
$CODY_{prä}$	0.48***	0.09	5.64
ri_{gesamt}	0.13	0.09	1.43
rs_{gesamt}	0.18*	0.08	2.19
Modell 2 ($R^2_{adjustiert}$ = 0.451; ΔR^2 = 0.033*)			

Anmerkung: Vollstandardisierte Lösung, daher Intercept in allen Modellen = 0. ri = random intercept im Lerntest, rs = random slope, ΔR^2 stets im Vergleich zum Baseline-Modell 0; *$p < 0.05$, **$p < .01$, *** $p < .001$

(n = 33) untersucht, wie die Entwicklung im Lerntest (*random slope*) mit den statusdiagnostischen Interventionseffekten zusammenhängt. Nach Jabrayilov et al. (2016) lassen sich sieben Zustände unterscheiden, die die Entwicklung von statusdiagnostischem Prä- zu Posttest qualifizieren. Die ersten drei bezeichnen eine Verbesserung in nur klinischer (n = 5), nur statistischer (n = 1) oder beiderlei (n = 14) Hinsicht und die folgenden vier eine entsprechende rein klinische (n = 0, hier per definitionem nicht möglich), rein statistische (n = 2) oder klinische und statistische (n = 0, hier per definitionem nicht möglich) Verschlechterung oder keine Veränderung (n = 11). Für alle drei Scorings finden sich deskriptiv betrachtet günstigere mittlere Lernverläufe (= positivere slopes) in der Gruppe mit Verbesserung (n = 20) im Vergleich zur Gruppe ohne Verbesserung (n = 13), wobei der Effekt nur für das Effizienzscoring statistisch vom Zufall unterscheidbar ist (d = 0.80, p = .04).

Diskussion

In diesem Beitrag wurde der computergestützte Lernverlaufstest „Goldmünzenjagd", der in ein Online-Training für Grundschulkinder mit Rechenschwierigkeiten eingebettet ist, vorgestellt. Da es sich bei der „Goldmünzenjagd" um ein computergestütztes Verfahren mit automatischer Administration und Auswertung handelt, ist die Durchführungs- und Auswertungsobjektivität hoch. Gleichzeitig ist eine unmittelbare Ergebnisrückmeldung an Eltern, Lehrkräfte oder Lerntherapeutinnen und –therapeuten möglich. Der Test wurde im Sinne des *robust indicator*-Ansatzes (Fuchs, 2004) konstruiert und bildet den Lernfortschritt im arithmetischen Faktenwissen (Addition und Subtraktion) sowie im Zahlenordnen ab. Für alle drei Inhaltsbereiche zeigten sich ähnliche Verlaufsmuster (vgl. Abbildung 2), wenn auch vergleichsweise höhere Geschwindigkeiten im Zahlenordnen und eine vergleichsweise geringere Präzision und Effizienz bei den Subtraktionsaufgaben.

Die beiden klassischen Scoring-Varianten, die Anzahl bearbeiteter Aufgaben und der Anteil korrekter Lösungen, wurden mit einer neuen HSHS-Scoringvariante, die Geschwindigkeit und Präzision auf Itemebene kombiniert und als Gewinn oder Verlust von Goldmünzen abbildet, verglichen. Zusammenfassend zeigen die Analysen, dass der beschriebene Lernverlaufstest ein reliables und valides Tool zur formativen Evaluation der Leistungsentwicklung von Grundschulkindern in basalen mathematischen Kompetenzbereichen ist. Von allen drei Scorings erwies sich das Münzscoring am reliabelsten, wobei die Splithalf-Reliabilitiäten ab dem zweiten Testzeitpunkt mit Werten über .90 am oberen Rand des von Klauer (2006) berichteten Bereichs liegen. Die Kriteriumsvaliditäten für die Scorings liegen mit .39-.51 etwas unterhalb des typischen Bereichs (Klauer, 2006). Als Kriterium wurde allerdings in unserer Studie kein klassischer Schulleistungstest, sondern ein Screening mit basisnumerischem und verglichen mit dem Lernverlaufstest stärker nicht-symbolischem Schwerpunkt (Kuhn et al., 2013) herangezogen. Dieses CODY-Screening ist genauso wie der Lernverlaufstest an die Intervention gekoppelt. Bemerkenswert ist, dass die Kriteriumsvaliditäten innerhalb der einzelnen Klassenstufen deutlich höher ausfallen als für die Gesamtstichprobe aller Klassenstufen (vgl. Tabelle 3). Dieser Befund kann dadurch erklärt werden, dass die Korrelation von normierten (aus dem CODY-Screening) und nicht-normierten Werten (aus dem Lernverlaufstest) zu konservativen Schätzungen psychometrischer Eigenschaften führt. Die Variation in den Lerntestwerten, die auf die Klassenstufe zurückgeht, bleibt in den Analysen mit den Rohwerten der Gesamtstichprobe entsprechend unkontrolliert. Dies scheint folglich auch bei einem nach dem *robust indicator*-Ansatz und somit per definitionem relativ curriculumsunabhängigen Verfahren eine Rolle zu spielen. Dabei ist jedoch zu beden-

ken, dass kognitive Effekte, die auf das Alter der Kinder zurückgehen und nicht mit dem Curriculum zusammenhängen (z. B. Effizienz der Informationsverarbeitung), mit der Klassenstufe konfundiert sind. Im Vergleich der Inhaltsbereiche fallen die psychometrischen Kennwerte für den Subtest Zahlenordnen am geringsten aus. Dieser Befund lässt sich durch den, relativ zu den Additions- und Subtraktionsaufgaben, heterogeneren Itempool erklären. Außerdem unterscheidet sich der Subtest Zahlenordnen im Antwortformat von den beiden Rechenfaktensubtests. Während die Ergebnisse der Additions- und Subtraktionsaufgaben über ein offenes Antwortformat eingetragen werden müssen, ist beim Zahlenordnen eine dichotome Auswahl zu treffen, die potentiell durch Ratetendenzen beeinflusst wird.

Um die Änderungssensitivität zu untersuchen, wurde die Prädiktivität der individuellen Ergebnisse in den Lernverlaufstests für die statusdiagnostische Entwicklung analysiert: Für alle drei Scorings zeigte sich eine inkrementelle Bedeutsamkeit individueller Lernverläufe über den Einfluss des statusdiagnostischen Ausgangsniveaus hinaus. Dies spricht dafür, dass der Lernverlaufstest den Anspruch der Änderungssensitivität erfüllt. Die Aussage wird durch den Vergleich zweier Teilgruppen unter den leistungsschwächsten der teilnehmenden Kinder (also der Hauptzielgruppe für das Training) unterstützt. Für die Teilgruppe, die sich in Folge des Trainings im statusdiagnostischen CODY-Test verbessert hatte, zeigten sich positivere Lernverläufe in der „Goldmünzenjagd" als für die Teilgruppe ohne Verbesserung im CODY-Posttest. Im Vergleich der Scorings zeigte sich, dass in der Geschwindigkeit und Effizienz (Münzen) vor allem die unterschiedlichen Lernverläufe einen Erklärungswert für die Leistung am Ende der Intervention hatten, während unter dem Präzisionsscoring eher das Ausgangsniveau im Lerntest prädiktiv für die statusdiagnostische Entwicklung war. Dieser Befund ist plausibel, wenn man ihn in Zusammenhang mit dem Konzept des *Overlearnings* (z. B. Klauer, 2006) betrachtet: Wenn bereits ein ausgeprägtes Präzisionsniveau erreicht ist, ist vor allem bei lernschwachen Kindern eine zusätzliche Verbesserung der Sicherheit und Geschwindigkeit anzustreben. Um diesen Lernprozess zu unterstützen, bietet das Münzscoring eine direkt ersichtliche Anreizstruktur, die schlechter Performanz aufgrund von Motivationsdefiziten vorbeugen sollte („Can't do"- statt „Won't do"-Assessment, vgl. Voß, 2016). Gleichzeitig soll diese Form der operanten Konditionierung zu einem schnellen, flüssigen Faktenabruf statt zum langsamen zählenden Rechen motivieren, und damit zu einer effizienteren Rechenstrategie, die für die weitere arithmetische Entwicklung bedeutsam ist, führen. Insgesamt zeigte sich, dass in der untersuchten Stichprobe eine effektivere Automatisierung einfacher Rechenfakten und Zahlenverarbeitung mit einem besseren Ansprechen auf die computerbasierte Förderung verbunden war. Die Ergebnisse reihen sich damit in die bisher überschaubare Befundlage ein, die belegt, dass selbst in höheren Klassenstufen die Effizienz mathematischen Faktenabrufs sowie ordinaler Zahlenverarbeitung (inkrementelle) Validität besitzt (z. B. Lyons et al., 2014; Nelson, Parker & Zaslofsky, 2016).

Limitationen

Neben den oben besprochenen Vorteilen impliziert der *robust indicator*-Ansatz zur Konstruktion von Lernverlaufsdiagnostika auch einige Limitationen, die nicht unerwähnt bleiben sollen. Dem großen Informationswert für Schülerinnen und Schüler im unteren Leistungsbereich steht ein eingeschränkter Nutzen für die Lernstandsevaluation auf Klassenebene entgegen. Insbesondere können über kompakte, simple Indikatoren (Anzahl korrekt gelesener Worte pro Zeit im Lesen, Effizienz beim Lösen von Additions- und Subtraktionsaufgaben im mathematischen Bereich etc.) keine übergeordneten Prozesse wie Strategiewissen erfasst werden. Aus der Leistungsentwicklung im

Verlauf der „Goldmünzenjagd“ lassen sich dementsprechend keine expliziten Aussagen hierüber ableiten. Implizit kann jedoch die zunehmende Geschwindigkeit und Effizienz als Indikator für einen verstärkten Gedächtnisabruf von Rechenfakten angesehen werden. Als eine weitere Limitation ist zu erwähnen, dass die hier berichteten Daten in einer Feldstudie gewonnen wurden, die einerseits eine hohe ökologische Validität bietet, bei der aber andererseits keine kontrollierten, standardisierten Durchführungsbedingungen hergestellt werden konnten. Weil die Evaluation des Trainings nicht im Vordergrund stand, wurde keine untrainierte Kontrollgruppe erhoben, an der die Interventionseffekte relativiert werden könnten. Die Analysen zur klinischen und statistischen Verbesserung sind deshalb explorativ zu verstehen und dürfen wegen der geringen statistischen Power nicht überinterpretiert werden.

Fuchs (2004) beschreibt drei Stadien der CBM-Forschung. In einem ersten Stadium geht es darum, neu entwickelte Verlaufsdiagnostika hinsichtlich ihrer psychometrischen Güte zu überprüfen. An dieser Stelle unterscheidet sich die Forschung nicht von der Evaluation statusdiagnostischer Tests. In einem zweiten Stadium geht es um den Nachweis, dass die Tests auch dazu geeignet sind, Lernentwicklungen abzubilden. Hier stellen sich beispielsweise die Forschungsfragen nach der Variabilität bei wiederholten Messungen, der Variabilität um typische Zuwachsraten sowie der Validität der Slopes als Maß für den Lernzuwachs in der entsprechenden Domäne. Für diese beiden Stadien des neu entwickelten Lerntests „Goldmünzenjagd“ liefert die hier berichtete Studie erste Evidenz. Daran sollte ein drittes Stadium anschließen, in dem die Implementation des neuen Instruments im Lehrkontext untersucht wird. Implementationsforschung aus diesem Stadium steht für den vorgestellten Lerntest noch aus.

Implikationen für Forschung und Praxis

Aus den Limitationen dieser Studie ergeben sich eine Reihe weiterführender Forschungsfragen: *Erstens* haben auf psychometrischer Ebene die Vergleiche der Kriteriumsvaliditäten gezeigt, dass bessere Kennwerte innerhalb von Klassenstufen als über Klassenstufen hinweg erreicht werden. Wie oben bereits diskutiert, spricht dies dafür, dass Berechnungen mit den Rohwerten der Lernverlaufstests zu konservativen Schätzungen der psychometrischen Parameter führen. Ein Forschungsdesiderat besteht deshalb in einer Normierungsstudie, die systematisch Schülerinnen und Schüler aus unterschiedlichen Altersgruppen einschließt. Dabei sollten auch weiteren Kriterien zur Validierung einbezogen werden, insbesondere curriculare Testverfahren in Abgrenzung zum hier eingesetzten Screening mit basisnumerischem Schwerpunkt. *Zweitens* wären die Aussagen zur Änderungssensitivität des Verfahrens belastbarer, wenn sie sich in Studien mit Kontrollgruppen replizieren ließen. Das bedeutet konkret, dass das hier verwendete Untersuchungsdesign mit einer Fördergruppe um eine unbehandelte Kontrollgruppe ergänzt werden sollte, um wie bei Klauer und Strathmann (2013) die Lernverläufe beider Gruppen vergleichen zu können. *Drittens* stehen Nachweise für die praktische Gültigkeit, Nutzung und Nützlichkeit des Verfahrens in weiteren als dem hier untersuchten privaten Kontext aus. Dies betrifft zum einen die Validierung im schulischen oder förderpädagogischen Bereich. In diesem Fall sollte die „Goldmünzenjagd“ während eines regulären Unterrichtszeitraums ohne das spezifische CODY-Training eingesetzt und die damit erfassten Lernverläufe ausgewertet werden. Als Validierungskriterium wären in diesem Setting statusdiagnostische Veränderungsmaße gleichermaßen wie die Urteile der pädagogischen Fachkräfte denkbar. Mit einer entsprechenden Studie ließe sich die Frage beantworten, ob der Lernver-

laufstest, der als interventionsbegleitendes Instrument konzipiert wurde, auch jenseits des Interventionskontextes informativ ist.

An solche Modellversuche knüpfen Fragestellungen der Implementationsforschung an, die beantwortet werden müssen, bevor eine flächendeckende Implementation denkbar ist (Hasselhorn, Köller, Maaz & Zimmer, 2014): Wie wird der relative Nutzen, die Komplexität und Durchführbarkeit dieser Art von Diagnostik eingeschätzt? Von welchen Merkmalen der Fachkräfte (z. B. Einstellungen), der Institution (z. B. technische Infrastruktur) oder des Umfeldes (z. B. implementationsbegleitende Fortbildungsangebote) ist dies abhängig? Welche Kriterien sind aussagekräftig, um den Implementationserfolg zu erfassen: die Geschwindigkeit, das Ausmaß oder die Tiefe der Verankerung? Welche Erfolgsmaße interessieren aufseiten der pädagogischen Fachkräfte oder der Schülerinnen und Schüler? Wie wirken sich schließlich die gemessenen Lernverläufe auf instruktionale Entscheidungen im Sinne adaptiver Förderung aus und wie werden Ausgangs- und Zieldaten verknüpft (Klauer, 2006)?

Schlussfolgerung

Ähnlich wie LEVUMI (Gebhardt, Diehl & Mühling, 2015) für den Bereich Lesen ist die „Goldmünzenjagd" als niedrigschwellig einsetzbares, internetbasiertes Verlaufsdiagnostikum für den basalen mathematischen Bereich gedacht. Durch die *robust indicator*-Konstruktion handelt es sich um kein im engeren Sinne curriculumsbasiertes Instrument, sondern um ein Verfahren, das über verschiedene Klassenstufen hinweg, insbesondere bei Kindern im unteren Leistungsbereich, eingesetzt werden kann. Eine Verknüpfung mit klassenstufenspezifischen Lehrplänen ist nicht vorgenommen worden: Die „Goldmünzenjagd" wurde explizit für rechenschwache Kinder mit dem Ziel konzipiert, einerseits änderungssensitiv zu sein, andererseits diese oft in ihren rechnerischen Fertigkeiten und ihrem Selbstbild stark benachteiligten Kinder zu motivieren. Die Besonderheiten der „Goldmünzenjagd" bestehen zum einen in der direkten Verzahnung mit einer Intervention und zum anderen in dem effizienzbezogenen HSHS-Scoring, das bei hier guter psychometrischer Qualität ein direktes Feedback über die Kombination von Präzision und Geschwindigkeit bietet. Für bestimmte diagnostische Fragestellungen kann es sich trotzdem anbieten, eine der beiden Komponenten – Präzision oder Geschwindigkeit – im Einzelnen zu betrachten. Während die Stichprobe dieser Studie den Verlaufstest begleitend zum Training zuhause durchgeführt hat, wäre zukünftig ebenso eine Implementation des Verfahrens in den institutionellen Förderkontext und inklusiven Unterricht denkbar. Solche Modellversuche sollten durch systematische Implementationsforschung begleitet werden.

Förderung

Diese Publikation ist im Rahmen eines durch das BMBF geförderten Projekts entstanden (Förderkennzeichen 01-GJ1302).

Literatur

Andersson, U. (2010). Skill development in different components of arithmetic and basic cognitive functions: Findings from a 3-year longitudinal study of children with different types of learning difficulties. *Journal of Educational Psychology, 102*, 115-134.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1-48.

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

Foegen, A., Jiban, C. & Deno, S. (2007). Progress monitoring measures in mathema-

tics: A review of the literature. *The Journal of Special Education*, *41*, 121-139.

Fuchs, L. S. (2004). The past. present. and future of curriculum-based measurement research. *School Psychology Review*, *33,* 188-193.

Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D. & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology*, *97*, 493-513.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Phillips, N. B. & Bentz, J. (1994). Classwide curriculum-based measurement: Helping general educators meet the challenge of student diversity. *Exceptional Children*, *60*, 518-537.

Fuchs, L. S., Powell, S. R., Seethaler, P. M., Cirino, P. T., Fletcher, J. M., Fuchs, D., … Zumeta, R. O. (2009). Remediating Number Combination and Word Problem Deficits Among Students With Mathematics Difficulties: A Randomized Control Trial. *Journal of Educational Psychology, 101,* 561–576.

Fuson, K. C., Wearne, D., Hiebert, J. C., Murray, H. G., Human, P. G., Olivier, A. I., Carpenter, T. & Fennema, E. (1997). Children's conceptual structures for multidigit numbers and methods of multidigit addition and subtraction. *Journal for Research in Mathematics Education*, *28*, 130-162.

Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities, 37,* 4–15.

Geary, D. C. (2013). Early foundations for mathematics learning and their relations to learning disabilities. *Current Directions in Psychological Science, 22*, 23-27.

Geary, D. C., Hoard, M. K. & Bailey, D. H. (2012). Fact retrieval deficits in low achieving children and children with mathematical learning disability, *Journal of Learning Disabilities*, *45*, 291–307.

Gebhardt, M., Diehl, K. & Mühling, A. (2015). Online-Lernverlaufsmessung für alle Schülerinnen und Schüler in inklusiven Klassen. *Zeitschrift für Heilpädagogik, 66*, 444-453.

Grömping, U. (2006). Relative importance for linear regression in R: the package relaimpo. *Journal of Statistical Software*, *17*, 1-27.

Hasselhorn, M., Köller, O., Maaz, K. & Zimmer, K. (2014). Implementation wirksamer Handlungskonzepte im Bildungsbereich als Forschungsaufgabe. *Psychologische Rundschau, 65,* 140-149.

Hosp, M. K., Hosp, J. L. & Howell, K. W. (2007). *The ABCs of CBM. A practical guide to curriculum-based measurement* (The Guilford practical intervention in the schools series). New York: Guilford Press.

Jabrayilov, R., Emons, W. H. & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Applied Psychological Measurement*, *40*, 559-572.

Jacobson, N. S. & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19.

Jansen, B. R. J., Louwerse, J., Straatemeier, M., Van der Ven, S. H. G., Klinkenberg, S. & Van der Maas, H. L. J. (2013). The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences, 24*, 190-197.

Jordan, N.C., Hanich, L.B. & Kaplan, D. (2003). Arithmetic fact mastery in young children: A longitudinal investigation. *Journal of Experimental Child Psychology, 85,* 103-119.

Kaufmann, L., Handl, P. & Thöny, B. (2003). Evaluation of a numeracy intervention program focusing on basic numerical knowledge and conceptual knowledge: A pilot study. *Journal of Learning Disabilities*, *36*, 564-573.

Klauer, K. J. (2006). Erfassung des Lernfortschritts durch curriculumbasierte Messung. *Heilpädagogische Forschung, 32*, 16-26.

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

Klauer, K. J. (2014). Formative Leistungsdiagnostik: Historischer Hintergrund und Weiterentwicklung zur Lernverlaufsdiagnostik. In M. Hasselhorn, U. Trautwein & W. Schneider (Hrsg.), *Lernverlaufsdiagnostik* (Vol. N.F. Band 12) (S. 1-17). Göttingen: Hogrefe.

Klauer, K. J. & Strathmann, A. M. (2013). Lernverlaufsdiagnostik Mathematik: Test auf Änderungssensibilität bei rechenschwachen Grundschülern. *Psychologie in Erziehung und Unterricht, 60*, 241-252.

Klinkenberg, S., Straatemeier, M. & Van der Maas, H. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*, 1813-1824.

Kuhn, J.-T. (2016). *Meister CODY: Computerbasiertes Trainingsprogramm für Grundschulkinder mit Rechenschwierigkeiten*. Beitrag auf dem 6. Frankfurter Forum (März 2016), Frankfurt am Main. Online verfügbar unter: https://www.testzentrale.de/veranstaltungen/frankfurter-forum/6-frankfurter-forum-2016.

Kuhn, J.-T. & Holling, H. (2014). Number sense or working memory? The effect of two computer-based trainings on mathematical skills in elementary school. *Advances in Cognitive Psychology, 10*, 59-67.

Kuhn, J.-T., Raddatz, J., Holling, H. & Dobel, C. (2013). Dyskalkulie vs. Rechenschwäche: Basisnumerische Verarbeitung in der Grundschule. *Lernen und Lernstörungen, 2*, 229-247.

Kuhn, J.-T., Schwenk, C., Strehle, L. M., Raddatz, J., Dobel, C. & Holling, H. (2017). *Evaluation of a computer-based training for enhancing arithmetic skills in math-disabled children*. Vortrag auf der 17. EARLI Conference (August/September 2017), Tampere.

Lyons, I. M. & Ansari, D. (2015). Numerical Order Processing in Children: From Reversing the Distance-Effect to Predicting Arithmetic. *Mind, Brain, and Education*, *9*, 207-221.

Lyons, I. M., Price, G. R., Vaessen, A., Blomert, L. & Ansari, D. (2014). Numerical predictors of arithmetic success in grades 1–6. *Developmental Science*, *17*, 714-726.

Maris, G. & Van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika, 77,* 615-633.

Nelson, P. M., Parker, D. C. & Zaslofsky, A. F. (2016). The relative value of growth in math fact skills across late elementary and middle school. *Assessment for Effective Intervention, 41*, 184-192.

Powell, S. R., Fuchs, L. S., Fuchs, D., Cirino, P. T. & Fletcher, J. M. (2009). Effects of Fact Retrieval Tutoring on Third-Grade Students with Math Difficulties with and without Reading Difficulties. *Learning Disabilities Research & Practice*, *24*, 1-11.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Wien. https://www.R-project.org/.

Raddatz, J., Kuhn, J. T., Holling, H., Moll, K. & Dobel, C. (2016). Comorbidity of arithmetic and reading disorder: Basic number processing and calculation in children with learning impairments. *Journal of Learning Disabilities, 50,* 298-308.

Rohwer, G. (2015). Bemerkungen zu einem Testverfahren für Lernfortschritte. *Journal for Educational Research Online*, *7*, 147-156.

Souvignier, E., Förster, N. & Salaschek, M. (2014). quop: Ein Ansatz internetbasierter Lernverlaufsdiagnostik mit Testkonzepten für Lesen und Mathematik. In M. Hasselhorn, U. Trautwein & W. Schneider (Hrsg.), *Lernverlaufsdiagnostik* (Vol. N.F. Band 12). Göttingen: Hogrefe.

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). *LVD-M 2-4. Lernverlaufsdiagnostik-Mathematik für zweite bis vierte Klassen (Hogrefe Schultests)*. Göttingen: Hogrefe.

Vanbinst, K., Ceulemans, E., Ghesquière, P. & De Smedt, B. (2015). Profiles of children's arithmetic fact development: A model-based clustering approach. *Journal of Experimental Child Psychology*, *133*, 29-46.

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.

Walter, J. (2010*). LDL. Lernfortschrittsdiagnostik Lesen. Ein curriculumsbasiertes Verfahren (Deutsch Schultests)*. Göttingen: Hogrefe.

Walter, J. (2013*). VSL. Verlaufsdiagnostik sinnerfassenden Lesens. (Hogrefe Schultests)*. Göttingen: Hogrefe.

Wißmann, J., Heine, A., Handl, P. & Jacobs, A. M. (2013). Förderung von Kindern mit isolierter Rechenschwäche und kombinierter Rechen-und Leseschwäche: Evaluation eines numerischen Förderprogramms für Grundschüler. *Lernen und Lernstörungen, 2,* 91-109.

Christin Schwenk, M.Sc.
Westfälische Wilhelms-Universität
Münster
Institut für Psychologie
Fliednerstraße 21
48149 Münster
christin.schwenk@uni-muenster.de

Erstmalig eingereicht: 28.02.2017
Überarbeitung eingereicht: 09.06.2017
Angenommen: 01.08.2017