Empirische Sonderpädagogik, 2015, Nr. 3, S. 206-222
ISSN 1869-4845 (Print) · ISSN 1869-4934 (Internet)

Lernverlaufsdiagnostik im Mathematikunterricht der zweiten Klasse: Raschanalysen und Empfehlungen zur Adaptation eines Testverfahrens für den Einsatz in inklusiven Klassen

Markus Gebhardt[1], Jörg-Henrik Heine[1], Nina Zeuch[2] & Natalie Förster[2]

[1] *TU München*
[2] *Westfälische Wilhelms-Universität Münster*

Zusammenfassung

Das Ziel der Lernverlaufsdiagnostik ist es, Lernverläufe von Schülerinnen und Schülern darzustellen. Lernverlaufsdiagnostik stellt insbesondere in leistungsheterogenen Klassen eine wichtige Informationsbasis für pädagogische Entscheidungen dar, indem für Schülerinnen und Schüler aller Leistungsniveaus rückgemeldet wird, ob der Unterricht angemessene Lernfortschritte bewirkt. Das internetbasierte System „quop" eröffnet die Möglichkeit, die Lernverläufe von einzelnen Schülerinnen und Schülern sowie von ganzen Klassen im Regelunterricht zu erfassen. Die hier bereitgestellten Testreihen wurden auf der Basis der klassischen Testtheorie konstruiert und sind jeweils für eine bestimmte Klassenstufe konzipiert. Für den Einsatz in sehr leistungsheterogenen Klassen wäre es wünschenswert, dass die Itemparameter der Aufgaben in den Tests bekannt sind und die Items auch über Klassenstufen hinweg verbunden werden können. Dafür ist es notwendig, dass der Test der probabilistischen Testtheorie entspricht. Ziel des vorliegenden Beitrags ist es, die dafür notwendigen Voraussetzungen im Hinblick auf die Eindimensionalität, Schwierigkeit und Testfairness der Testreihe zur Lernverlaufsdiagnostik im Mathematikunterricht in zweiten Klassen auf Basis des Raschmodells zu prüfen. Die Analysen der Daten von 414 Schülerinnen und Schülern aus 19 Klassen zeigen, dass anhand der Testreihe die Leistungsentwicklung im Mathematikunterricht auf den beiden Dimensionen Vorläuferkompetenzen und curriculare Kompetenzen reliabel erfasst werden können. Eine mögliche Weiterentwicklung der beiden Subskalen für die Nutzung in inklusiven Klassen wird diskutiert.

Schlagwörter: Lernverlaufsdiagnostik, Mathematik, 2. Jahrgangstufe, Grundschule, Längsschnittstudie, Item Response Theorie

Learning progress assessment in mathematic in second grade: Rasch analysis and recommendations for adaptation of a test instrument for inclusive classrooms

Abstract

Learning progress assessment aims to monitor student learning growth. Especially in heterogeneous classrooms, learning progress assessment provides important information that can be used in the instructional decision-making process by monitoring whether the instruction is ef-

fective for students at all skill levels. The web-based system „quop“ provides an opportunity to monitor the learning progress of individual students and also of whole classrooms in general education. The test series available on the quop-platform were designed on the basis of classical test theory for a specific grade level. However, it would be helpful to analyze item parameters of the tests in order to use them across grades in highly heterogeneous classrooms. Therefore the tests should meet the requirements of item response theory. The aim of this paper is to examine the necessary conditions in terms of one-dimensionality, difficulty and test fairness of the test series for learning progress assessment in mathematics in second grade based on the Rasch model. Data analysis of 414 students from 19 classes shows that the tests reliably assess the development in mathematics on the two dimensions precursor competencies and curricular skills. Further development of the two subscales for use in inclusive classrooms is discussed.

Keywords: learning progress assessment, mathematics, second grade, primary school, longitudinal design, Item Response Theory

Lernverlaufsdiagnostik in der inklusiven Schule

Sowohl die Lernausgangslagen als auch Lernentwicklungen von Schülerinnen und Schülern sind heterogen (Prenzel et al., 2006). Insbesondere in inklusiven Schulklassen, in denen Schülerinnen und Schüler mit und ohne sonderpädagogischem Förderbedarf (SPF) gemeinsam unterrichtet werden, stellt die Heterogenität der schulischen Leistungen eine besondere Herausforderung für Lehrkräfte dar. Um die positive Entwicklung der Lernenden in inklusiven Klassen sicherzustellen, wird von der Schulaufsicht gefordert, dass eine (formative) Diagnostik mit dem Ziel der Unterrichtsanpassung durchgeführt und daraus abgeleitete Förderziele in einem Förderplan dokumentiert werden (Bundschuh, 2010; Heimlich, Lotter & März, 2005). Das empfohlene Vorgehen ist hier, dass standardisierte Schulleistungstests einmalig verwendet und basierend auf den Ergebnissen Fördermaßnahmen für das nächste Schuljahr abgeleitet werden. Die Schulleistungstests sind meistens für bestimmte Klassenstufen normiert und liefern somit der Lehrkraft Informationen über den Lernstand einer Schülerin oder eines Schülers im Vergleich zum Leistungsstand aller Schülerinnen und Schüler der Klassenstufe. Bei derartigen Testungen befinden sich Schülerinnen und Schüler mit einem SPF häufig am unteren Rand der Normierung (Gebhardt, Schwab, Krammer & Gasteiger-Klicpera, 2012; Wocken & Gröhlich, 2009). Für eine formative Diagnostik sind diese Tests nur eingeschränkt praktikabel, da die Durchführung häufig bis zu einer Schulstunde oder darüber hinaus dauert und oft keine ausreichende Anzahl an parallelen Testversionen vorliegt.

Neuere Unterrichtskonzepte wie zum Beispiel der Ansatz Response to Intervention (RTI; Huber & Grosche, 2012) basieren maßgeblich auf dem Ansatz formativer Diagnostik, um individuelle Rückmeldungen über Leistungsstände und -entwicklungen zu erhalten (Blumenthal, Kuhlmann & Hartke, 2014). Ein möglicher Weg, Lernverläufe von Schülerinnen und Schülern zu erfassen, ist die Durchführung von Curriculum Based Measurements (CBM; Deno, 2003). In kurzen zeitlichen Abständen (z. B. wöchentlich) werden ökonomische Paralleltests durchgeführt. Durch die Darstellung des Lernverlaufs erhalten die Lehrkräfte ein Feedback über die Effektivität ihres Unterrichts und können Instruktionen für die gesamte Klasse oder einzelne Schülerinnen und Schüler anpassen. Evaluationsstudien zeigen überwiegend positive Effekte der Lernverlaufsdiagnostik, insbesondere auch bei Schülerinnen und Schülern mit schulischen Schwierigkeiten und mit SPF (Stecker, Fuchs & Fuchs, 2005). Erste deutsche Studien zur Wirksamkeit der Lernverlaufsdiagnostik stützen diese Befunde und zeigen darü-

ber hinaus, dass die Bereitstellung diagnostischer Information über Lernverläufe über eine statusdiagnostische Information hinaus zu höheren Lernzuwächsen bei den Schülerinnen und Schülern führt (Förster & Souvignier, 2014, 2015; Souvignier & Förster, 2011). Für den deutschsprachigen Raum existieren erste Instrumente zur Erfassung der Entwicklung des (verstehenden) Lesens (Walter, 2009, 2013) und der Mathematik (Strathmann & Klauer, 2012). Die Anforderungen an diese Form der Leistungsmessung sind insbesondere im Hinblick auf die Ökonomie der Tests hoch. Die einzelnen Tests müssen von kurzer Dauer und leicht in den Unterrichtsalltag integrierbar sein (Müller & Hartmann, 2009; Souvignier, Förster & Salaschek, 2014). Darüber hinaus sollten Lehrkräfte anhand der Ergebnisse gezielt Fördermaßnahmen ableiten können. Neben der Durchführungsökonomie spielen die Auswertungsökonomie sowie die Form der Ergebnisdokumentation eine entscheidende Rolle. Ein vielversprechender Zugang ist hier die Testung am Computer. Die Schülerinnen und Schüler erhalten unabhängig von der Lehrkraft standardisierte Testinstruktionen und können in ihrem eigenen Tempo den Test bearbeiten. Am Computer ist das Verteilen von verschiedenen Paralleltests, das Erfassen der Bearbeitungszeit sowie die Auswertung und Ergebnisdarstellung der Tests automatisiert möglich. Umgesetzt wurde dies beispielsweise in der internetbasierten Lernverlaufsdiagnostikplattform „quop" (Souvignier et al., 2014). Die Internetplattform bietet Testreihen zur Lernverlaufsdiagnostik für verschiedene Klassenstufen in der Grundschule in den Bereichen Lesen und Mathematik an, bei denen Schülerinnen und Schüler kurze Tests am Computer bearbeiten. Unmittelbar im Anschluss erhalten die Lehrkräfte schüler- und klassenbezogene Ergebnisdarstellungen sowie Vergleichswerte und Hilfestellungen zur Interpretation. Die bislang über die Plattform quop verfügbaren Testreihen wurden für den Einsatz im Regelunterricht konstruiert und überprüft (Souvignier et al., 2014). Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf wurden bis jetzt nicht explizit in quop berücksichtigt. Um in inklusiven Klassen alle Schülerinnen und Schüler testen zu können, müssten Lehrkräfte aktuell für Kinder mit SPF entsprechend des Lehrplanes eine andere Testreihe einer niedrigeren Klassenstufe auswählen, um leichtere Aufgaben und entsprechende Vergleichswerte zu erhalten. Ein direkter Vergleich mit Mitschülerinnen und Mitschülern ohne SPF ist dadurch nicht möglich. Wünschenswert für den Einsatz in inklusiven Klassen wäre, dass sich die Lernentwicklungen von Kindern über Klassenstufen hinweg auf derselben Skala dokumentieren ließen. Dafür geeignete Tests müssten auf einem Itempool basieren, deren Itemparameter über verschiedenen Klassenstufen hinweg verbunden sind. Basierend auf einem solchen Itempool wäre theoretisch auch eine adaptive Testung am Computer möglich, bei welcher anhand der Eingaben der Schülerin oder des Schülers die adäquate Aufgabe gezogen wird. Eine notwendige Voraussetzung für das adaptive Testen ist, dass ein nach den Prinzipien der probabilistischen Testtheorie kalibrierter Itempool vorliegt (z.B. Frey, 2012).

Testkonstruktion

Fuchs (2004) unterscheidet zwei Möglichkeiten, wie curriculumbasierte Tests konstruiert werden können: über robuste Indikatoren oder Curriculum Sampling. Beim ersten Ansatz werden Typen von Aufgaben gesucht, die die geforderte Kompetenz möglichst gut repräsentieren und hoch mit relevanten Leistungen korrelieren. Für die Lesekompetenz hat sich beispielsweise das Laute Lesen für eine Minute als ein robuster Indikator herausgestellt (Reschly, Busch, Betts, Deno & Long, 2009; Wayman, Wallace, Wiley, Ticha & Espin, 2007). Beim Curriculum Sampling hingegen werden die am Schuljahresende geforderten Kompetenzen in Teilmengen unterteilt. Jeder Paralleltest enthält dann dieselbe Anzahl der vorab definierten Aufgabentypen (Voß & Hartke, 2014). Un-

abhängig vom Konstruktionsansatz besteht eine notwendige Voraussetzung zur Veränderungsmessung darin, dass auch bei wiederholten Messungen stets dasselbe homogene Konstrukt gemessen wird, jeder Paralleltest die gleiche Testschwierigkeit hat und die Tests änderungssensibel sind (Klauer, 2014). Überprüfen kann man diese Voraussetzungen durch die Item Response Modelle aus der probabilistischen Testtheorie (IRT), welche von Wilbert und Linnemann (2011) explizit für die Skalierung von Lernverlaufsdiagnostik vorgeschlagen wird. Nach dieser Theorie wird eine latente Personeneigenschaft (Fähigkeit) bei der Auswertung der Tests angenommen, welche über Modellparameter modelliert wird. Dies geschieht einerseits durch die Ausprägung der Person auf der latenten Eigenschaft (Personenparameter) und andererseits anhand der Schwierigkeit der Aufgabe (Itemparameter). Die Wahrscheinlichkeit der Lösung einer Testaufgabe steht mit den beiden Parametern in einer psychologisch plausiblen probabilistischen Beziehung (Rost, 2004). Hierbei ist eine notwendige Voraussetzung zur Modellgültigkeit, dass die Eindimensionalität der Skala und die stichprobeninvariante Anordnung der Items nach ihrer Schwierigkeit gegeben sind. Erweist sich eine Skala in diesem Sinne als eindimensional und haben alle Items dieselbe Trennschärfe, liegt das Raschmodell vor. Erst wenn das Raschmodell gilt, sagt der Summenwert der Rohwerte alles über das Antwortverhalten der getesteten Personen und damit über deren latente Eigenschaft aus. Damit die Veränderung der Summenwerte auf eine Veränderung der Kompetenzen zurückgeführt werden kann, müssen die einzelnen eingesetzten Tests nicht nur dasselbe Konstrukt erfassen, sondern auch homogene Testschwierigkeiten besitzen (Klauer, 2014), was sowohl für die Konstruktion als auch die empirische Überprüfung der Paralleltests eine Herausforderung darstellt. Im Hinblick auf die Testfairness ist eine weitere Anforderung an die Tests, dass sie für verschiedene Personengruppen (unter Konstanthaltung der Personenparameter) gleich schwer sein sollen. Wenn darüber hinaus alle Items die gleiche Schwierigkeitsrangfolge bei unterschiedlichen Personenfähigkeiten besitzen, ist der Test fair. Für die Testkonstruktion zur Anwendung in der Lernverlaufsdiagnostik schlagen Wilbert und Linnemann (2011) vor, dass zuerst auf Basis der KTT die Trennschärfen und die Retest-Reliabilität der einzelnen Tests bestimmt werden und danach die Eindimensionalität nachgewiesen wird, bevor abschließend die Itemparameter der einzelnen Paralleltests sowie die Testfairness für relevante Subgruppen bestimmt werden.

Fragestellungen

Ziel des Beitrages ist es, die Testreihe zur Lernverlaufsdiagnostik im Fach Mathematik für Zweitklässler (Salaschek & Souvignier, 2014) anhand dieser Kriterien zu überprüfen. Zudem sollen basierend auf den Analysen Empfehlungen zur Adaption der Testreihe abgeleitet werden, um auch in inklusiven Klassen die Leistungsentwicklungen von Schülerinnen und Schüler erfassen zu können. Für den Mathematiktest ist die Gültigkeit nach klassischer Testtheorie nachgewiesen (ebd.). In diesem Beitrag erfolgt eine Re-Analyse der Daten nach dem Raschmodell (Rasch, 1960).

In der vorliegenden Untersuchung wurde geprüft, inwiefern sich die zwei bei der Testkonstruktion zugrunde gelegten Dimensionen Vorläuferkompetenzen und curriculare Aufgaben empirisch bestätigen lassen. Darüber hinaus wurden die Itemparameter der einzelnen Tests bestimmt.

Im Hinblick auf die Testfairness wurde untersucht, ob die Tests für unterschiedlich leistungsstarke Schülerinnen und Schüler sowie für Mädchen und Jungen messinvariant sind. Die Messinvarianz über Leistungsgruppen und Geschlechter ist eine Voraussetzung dafür, Testwerte des Instrumentes verschiedener Gruppen zu vergleichen. Abschließend wurde untersucht, inwiefern sich mit dem Instrument Lernverläufe von Schülerinnen und Schülern erfassen lassen.

Methode

Stichprobe

Insgesamt nahmen 414 Schülerinnen und Schüler (212 männlich) aus 19 zweiten Klassen an der Untersuchung teil. Der überwiegende Teil der Schülerinnen und Schüler (83%) sprach Deutsch oder Deutsch und eine andere Sprache zu Hause. Zu Beginn der Untersuchung waren die Schülerinnen und Schüler im Schnitt 7.60 Jahre alt (SD = 0.57).

Design und Instrument

In einem dreiwöchigen Rhythmus bearbeiteten die Schülerinnen und Schüler zwischen Oktober und Mai insgesamt acht Tests am Computer. Vier Paralleltests (Version A, B, C und D) wurden dazu jeweils zweimal verwendet, so dass jeweils zu Messzeitpunkt (MZP) 1 und MZP 5 (Version A), MZP 2 und MZP 6 (Version B), MZP 3 und MZP 7 (Version C) sowie MZP 4 und MZP 8 (Version D) identische Tests bearbeitet wurden. Jeder dieser Mathematiktests besteht aus den Dimensionen *Vorläuferkompetenzen* (24 Aufgaben) und *curricularen Aufgaben* (28 Aufgaben). Die Testkonstruktion ist bei Salaschek und Souvignier (2014) beschrieben und baut im Wesentlichen auf dem Modell von Krajewski (Krajewski & Schneider, 2009) auf. Es wird demnach angenommen, dass Vorläuferkompetenzen basale mathematische Kompetenzen darstellen (Zahlenverständnis, Mengenverständnis, Zahlen bestimmten Mengen oder Positionen auf Zahlenstrahlen zuordnen etc.; siehe auch Berch, 2005, im Überblick) und damit Voraussetzung für die adäquate Entwicklung von weiteren mathematischen Kompetenzen sind (z.B. Rechenfertigkeiten; siehe auch Krajewski & Schneider, 2009). Da anzunehmen ist, dass auch in der zweiten Klassenstufe noch nicht alle Schülerinnen und Schüler eine vollständige Beherrschung von Vorläuferkompetenzen aufweisen, ist es aufgrund ihrer Bedeutung für die Entwicklung weiterer mathematischer Kompetenzen wichtig, diese auch in der zweiten Klasse zu erfassen (vgl. auch Krajewski & Schneider, 2009).

Um die Vorläuferkompetenzen möglichst breit abzubilden, werden sie anhand unterschiedlicher Aufgabentypen erfasst: Zahlen bis 1000 erkennen (8 Items), Zahlenvergleich von zwei Geldbeträgen bis 100 € (6 Items), Zahlenstrahl bis 100 (6 Items) und Spiegelachsen (4 Items). Auch die curricularen Aufgaben umfassen verschiedene Aufgabentypen: Rechenaufgaben im Addieren (4 Items), Subtrahieren (4 Items) und Multiplizieren (4 Items), wie auch die Aufgaben Verdoppeln (6 Items), Halbieren (6 Items) und Ergänzen auf 100 (4 Items).

Die interne Konsistenz der Tests lag bei den acht Messzeitpunkten jeweils zwischen r_{α} =.87 und r_{α} =.96. Die Retest - Reliabilität lag zwischen r_{tt} = .81 und r_{tt} = .87 und bei einem Abstand von 12 Wochen bei r_{tt} = .77. In den Analysen von Souvignier et al. (2014) wurde mittels MANOVA ermittelt, dass es bei fast allen Testzeitpunkten signifikante Lernerfolge gab und die Tests somit sensitiv für Veränderungen sind. Die zwei Dimensionen *Vorläuferkompetenzen* und *curriculare Aufgaben* wurden mittels konfirmatorischer Faktorenanalyse für alle acht Zeitpunkte als akzeptabel nachgewiesen (Salaschek & Souvignier, 2014).

Statistische Analyse

Die Reanalysen werden mit dem Statistikprogramm R (R Core Team, 2013) und den Paketen TAM (Kiefer, Robitzsch & Wu, 2014) und pairwise (Heine, 2014) durchgeführt. Zur Schätzung der Modellparameter im Rahmen der Item Response Theorie werden in der empirischen Bildungsforschung üblicherweise entweder die Conditional-Maximum-Likelihood Schätzung (CML), die Marginal-Maximum- Likelihood -Schätzung (MML) oder die Joint-Maximum Likelihood (JML) eingesetzt (Heine, Sälzer, Borchert, Siberns & Mang, 2013). Neben diesen drei

standardmäßig eingesetzten Schätzverfahren besteht noch eine weitere Methode zur expliziten Berechnung der Itemparameter im Raschmodell nach der Methode des paarweisen Itemvergleichs (Choppin, 1968, 1985; Rasch, 1960; Rost, 2004; Wright & Masters, 1982). Diese Methode eignet sich insbesondere zur Bestimmung der stichprobeninvarianten Itemparameter für die Kalibrierung eines gegebenen Itempools (Choppin, 1968).

Mehrdimensionale Modelle werden üblicherweise mit der MML-Methode, welche in TAM implementiert ist, geschätzt (z.B. PISA 2012; Heine et al., 2013). Die Überprüfung der Dimensionalität des vorliegenden Testmaterials wird daher mit diesem Paket überprüft. Dabei werden zunächst für alle Testzeitpunkte alle Items einer eindimensionalen Skalierung und anschießend einer zweidimensionalen Skalierung unterzogen, wobei bei letzterer die Zuordnung der Items zu den Dimensionen theoriegeleitet vorgenommen wurde. Die Passung beider Modelle wird im Anschluss anhand informationstheoretischer Kriterien verglichen. Sollte eine mehrdimensionale Struktur innerhalb des Testmaterials gefunden werden, werden in den weiteren Analysen die einzelnen Skalen auf Raschhomogenität mittels eindimensionaler Skalierung untersucht und ausgewertet (Wilbert, 2014). Hierfür wird der Pairwise Schätzer gewählt, da dieser (im Gegensatz zu den o.g. weiteren Schätzern) auch bei fehlenden Daten und kleinen Stichproben robust schätzt (Heine & Tarnai, 2013, 2015; Wright & Masters, 1982). Nach Analyse der Itemparameter werden die Personenparameter mittels der Weighted-Maximum-Likelihood-Methode (WLE; Warm, 1989) geschätzt. Etwaige lokale Modellverletzungen werden über Item-Fit Statistik basierend auf den standardisierten Residuen der eindimensionalen Skalierung untersucht (Wright & Masters, 1982). Darüber hinaus werden für die dichotomen Items jeweils die punktbiserialen Korrelationen mit dem Skalenwert (WLE-Schätzer) berichtet.

Ergebnisse

Überprüfung der Dimensionalität der Testreihe

Für die Untersuchung der Dimensionalität des Testmaterials wird aufgrund der großen Itemzahl von 52 Aufgaben pro Testzeitpunkt das Bayesian Information Criterion (BIC; Schwarz, 1978) für die Bestimmung der Anzahl der Dimensionen gewählt (Rost, 2004). Der BIC ist für die einzelnen Modelle in der Tabelle 1 abgebildet. Zu erkennen ist, dass der BIC für ein Modell mit zwei Dimensionen niedriger ist. Somit beschreibt das zweidimensionale Modell die Daten besser als das eindimensionale Modell. Dieses Ergebnis steht im Einklang mit dem Ziel, einen zweidimensionalen Test zu entwickeln und stützt den mittels konfirmatorischer Faktorenanalyse gefunden Befund der Zweidimensionalität (Salaschek & Souvignier, 2014). Für die nachfolgenden Analysen wird daher diese Struktur beibehalten.

Überprüfung der Parallelität

Für jede Dimension wird eine eindimensionale Skalierung nach dem Raschmodell mittels der Methode der expliziten Parameterberechnung des paarweisen Vergleichs für jeden Messzeitpunkt separat durchgeführt.

Tabelle 1: BIC der 1PL Testmodelle für jeden Testzeitpunkt

	MZP 1	MZP 2	MZP 3	MZP 4	MZP 5	MZP 6	MZP 7	MZP 8
1 Dim	23648	23205	22393	21705	20361	17968	19715	18071
2 Dim	23565	23070	22270	21534	20241	17578	19573	17869

Anmerkungen. Dim = Dimensionen; MZP = Messzeitpunkt.

Durch die vier Paralleltests sind die zwei Tests zu den Zeitpunkten MZP 1 und MZP 5, MZP 2 und MZP 6, MZP 3 und MZP 7 sowie MZP 4 und MZP 8 jeweils identisch.

Die Theorie des Raschmodells postuliert als Voraussetzung der Modellgültigkeit, neben anderen Kriterien, die Stichprobeninvarianz der Itemparameter. Diese Stichprobeninvarianz bezieht sich einerseits auf unterschiedliche Substichproben wie z.B. das Kompetenzniveau oder das Geschlecht, kann aber auch bei konstanter Personenstichprobe auf zu unterschiedlichen Messzeitpunkten erhobenes Testmaterial bezogen werden. Insbesondere für die Messung des Lernverlaufs ist es daher vorteilhaft, wenn sich die Itemparameter als invariant über die Messzeitpunkte bzw. als konstant erweisen. Die relativen (summennormierten) Schwierigkeitsparameter einer Reihe von Items untereinander sollten sich also sowohl bei Kalibrierung der Items zu Messzeitpunkt 1 als auch bei Kalibrierung zu Messzeitpunkt 2 (oder anderen Messzeitpunkten) invariant ergeben. Aufgrund des Lernzuwachses der Personen ist dagegen zu erwarten, dass sich die Personenparameter über die Zeit verändern. Darüber hinaus ist zu erwarten, dass in den Paralleltests die gleichen Items die gleichen Itemparameter aufweisen.

Abbildung 2 zeigt beispielhaft für die Testzeitpunkte MZP 1 und MZP 5 (Parallelversion A) sowie MZP 2 und MZP 6 (Parallelversion B) den grafischen Modelltest zwischen jeweils zwei Messzeitpunkten. Die jeweils an der X- und Y-Achse abgetragenen Itemparameter der beiden Messzeitpunkte verlaufen entlang der Winkelhalbierenden, wobei die 95% Konfidenzintervalle der Itemparameter-Punktschätzer als Ellipsen dargestellt sind. Dieser Verlauf der Itemparameter legt einerseits die Modellgeltung nahe und belegt, dass die Items zu jeweils beiden Messzeitpunkten (annähernd) gleiche Schwierigkeiten aufweisen. Dieses Er-

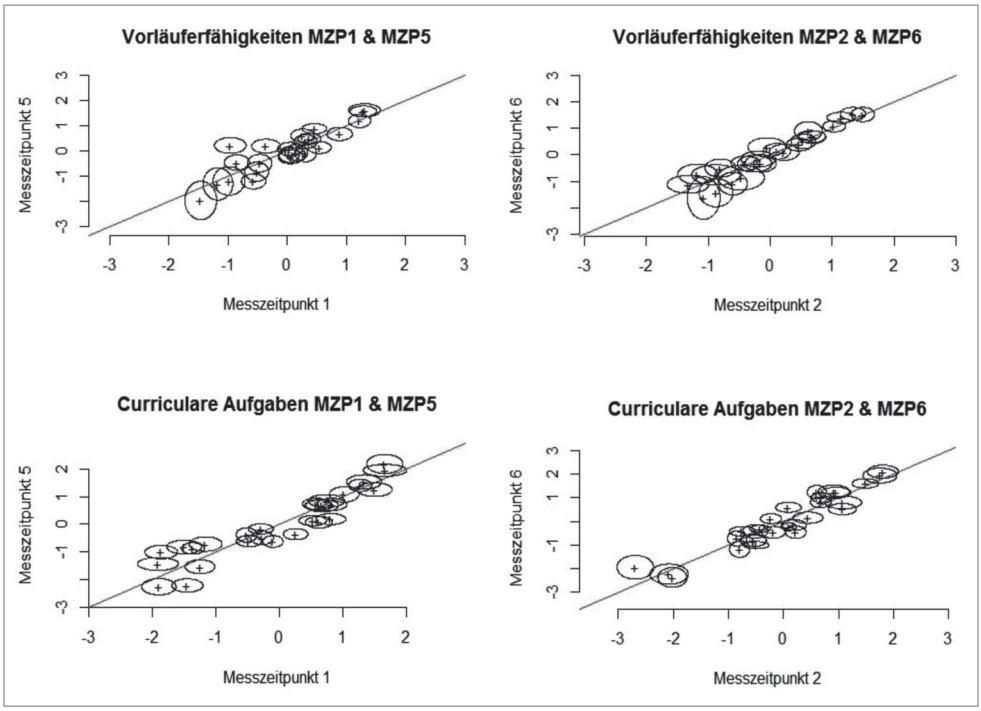

Abbildung 1: Analysen zur Messinvarianz nach Messzeitpunkten in den Dimensionen Vorläuferkompetenzen (oben) und Curriculare Aufgaben (unten)

gebnis gilt auch für die aus Platzgründen nicht dargestellten Messzeitpunkte 3 und 7 (Parallelversion C) sowie 4 und 8 (Parallelversion D).

Da die Paralleltests nicht zum selben Messzeitpunkt gemessen wurden und darüber hinaus auch nicht auf der Ebene der Items verlinkt sind, können die Itemparameter der verschiedenen Testversionen nicht auf einer gemeinsamen Skala miteinander verglichen werden. Daraus ergibt sich, dass auch die Lernerfolge der Schülerinnen und Schüler z. B. zwischen MZP1 und MZP5 (Parallelversion A) nicht mit denen zwischen MZP2 und MZP6 (Parallelversion B) verglichen werden können. Eine gemeinsame Schätzung eines Modells der Paralleltests ist somit nicht möglich. Daher wird nachfolgend nur der erste Paralleltest näher beschrieben. Für diesen Test sind zum ersten Testzeitpunkt noch keine Lerneffekte möglich.

Bestimmung der Itemparameter

Die Parallelversion A wurde zu MZP1 und MZP5 verwendet. Die Analysen zeigen weitgehend invariante Itemparameter der beiden Messzeitpunkte (MZP1 und MZP5). Das bedeutet, dass die Aufgaben in den identischen Tests zu MZP1 und MZP5 weitgehend die gleiche Schwierigkeitsrangfolge zeigen. Daher können die Itemparameter MZP1 zur Schätzung der Personenparameter zu MZP5 herangezogen werden, wodurch sich der Lernerfolg der Schülerinnen und Schüler zwischen den beiden Messzeitpunkten bestimmen lässt. Nachfolgend werden daher MZP1 und MZP5 ausgewertet, wobei zur Bestimmung der Personenfähigkeit zu den beiden Messzeitpunkten jeweils die Itemparameter von MZP1 verwendet werden. In Tabelle 2 und 3 sind die Itemparameter mit der Lösungswahrscheinlichkeit $p = .5$ in der Logit-Metrik dargestellt. Da anzunehmen ist, dass einzelne Facetten der mathematischen Kompetenz aufeinander aufbauen (wenn auch nicht streng linear; vgl. Krajewski & Schneider, 2009) sollten Aufgaben vom Typ „Zahl erkennen“ insgesamt früher beherrscht werden als Aufgaben vom Typ „Zahlenvergleich“. In Tabelle 2 erkennt man, dass der Subtest „Zahl erkennen“ insgesamt wie erwartet leichter ist als der Subtest „Zahlenvergleich“, einzelne Items aber nicht diesem Gesamtmuster folgen.

Überprüfung der Infit- und Outfit-Koeffizienten

Neben der oben bereits diskutierten Stichprobeninvarianz der Itemparameter als Indikator für die Globale Modellpassung stellt sich bezüglich der Geltung des Raschmodells stets auch die Frage nach lokalen Modellverletzungen (d. h. Verletzungen auf der Itemebene). Dafür können die Maße Infit und Outfit herangezogen werden. Der Outfit ist die Summe der quadrierten standardisierten Residuen und ist sensitiv für Raten und Flüchtigkeitsfehler (Outliers). Der Infit ist im Gegensatz zum Outfit nach Informationen gewichtet und zeigt Verzerrungen im Sample (wie z. B. Guttman Pattern). Die Infit- und Outfit-Koeffizienten sowie der punktbiserialen Korrelationskoeffizienten (vgl. Tabelle 2 und 3) deuten darauf hin, dass die meisten Items der beiden Dimensionen eine gute Passung bezüglich der Annahmen des Raschmodells aufweisen (Wright & Masters, 1982). Die in Tabelle 2 und 3 fett markierten Infit- und Outfit-Koeffizienten kennzeichnen die Items, deren Mean Square Fit-Werte signifikant von ihrem Erwartungswert 1 abweichen. Bei einem Wert unter 1 sind die Daten vorhersagbarer als das Raschmodell annimmt. Dies bedeutet, dass diese Items eher zu trennscharf sind und sich insgesamt die Antwortmuster im Rahmen der IRT Modellierung eher dem Guttman-Pattern annähern. Dieser Befund ist aber im Hinblick auf die intendierte summative Verrechnung der Schülerantworten zu einem Skalenwert und insbesondere unter Berücksichtigung der Länge der Skalen als unproblematisch einzustufen und bei Leistungstests nicht ungewöhn-

lich. Problematisch können dagegen Items mit einem signifikant über 1 liegenden Wert sein, welche einen erheblichen Underfit aufweisen. Dies kann auf zufälliges Antworten (Raten) oder aber auch Decken- bzw. Boden-Effekte im Hinblick auf das ‚Targeting' von Personenfähigkeit und Itemparameter hinweisen (Linacre, 2002). In Tabelle 2 und 3 liegen die vorgefundenen Infit- und Outfit-Werte zum MZP1 zwischen 0.5 und 1.5. In diesem Wertebereich kann davon ausgegangen werden, dass die Messung der mathematischen Kompetenz der Schülerinnen und Schüler nicht negativ beeinflusst

Tabelle 2: Itemparameter und punktbiseriale Korrelation für die Dimension Vorläufer zum ersten Messzeitpunkt in Paralleltestversion A und die Infit- und Outfit-Koeffizienten zu den Messzeitpunkten eins und fünf

Dimension Vorläufer							
MZP 1				MZP 1		MZP 5	
Item	**Subtest**	**Item-parameter**	**Punkt-biseriale Korrelation**	**OUTFIT MSQ**	**INFIT MSQ**	**OUTFIT MSQ**	**INFIT MSQ**
1	Zahl erkennen	-0.52	.28	1.08	1.06	0.86	0.84
2	Zahl erkennen	-1.46	.34	0.77	0.92	0.68	**0.65***
3	Zahl erkennen	-1.17	.27	1.13	0.95	1.04	0.82
4	Zahl erkennen	-0.85	.33	0.95	0.97	1.36	**1.27***
5	Zahl erkennen	0.05	.45	0.93	0.98	0.82	0.94
6	Zahl erkennen	1.22	.51	0.96	0.97	0.94	0.93
7	Zahl erkennen	-0.36	.47	0.79	0.93	1.09	**1.27***
8	Zahl erkennen	0.26	.52	0.85	0.89*	0.98	1.01
9	Zahlvergleich	0.09	.45	0.94	0.96	0.99	0.94
10	Zahlvergleich	0.16	.40	1.02	1.03	0.85	0.86*
11	Zahlvergleich	0.03	.48	0.90	0.92	0.76	0.87
12	Zahlvergleich	0.46	.34	1.17*	1.14*	1.07	1.14*
13	Zahlvergleich	1.31	.37	**1.24***	1.11*	1.02	1.01
14	Zahlvergleich	1.28	.50	0.97	0.97	0.91	0.94
15	Zahlenstrahl	-0.96	.26	0.99	1.03	**2.49***	**2.14***
16	Zahlenstrahl	0.05	.46	0.94	0.94	0.83	0.87
17	Zahlenstrahl	0.31	.42	1.05	1.01	1.06	1.05
18	Zahlenstrahl	-0.47	.41	0.91	0.98	0.82	0.98
19	Zahlenstrahl	0.89	.36	1.17*	1.13*	1.13	1.11
20	Zahlenstrahl	0.38	.50	0.92	0.93	0.90	1.00
49	Spiegelachse	0.55	.42	1.02	1.04	0.90	0.92
50	Spiegelachse	-0.99	.29	1.17	0.99	0.91	0.84
51	Spiegelachse	-0.57	.34	1.00	1.02	0.59	**0.65***
52	Spiegelachse	0.31	.34	1.14	1.12*	0.99	0.95

Anmerkungen. MZP = Messzeitpunkt. *signifikante Abweichung ($p < .05$) vom Mean Square Fit (MSQ); fett markiert sind auffällige signifikante Werte unter 0.8 und über 1.2

Tabelle 3: Itemparameter und punkt-biseriale Korrelation für die Dimension curriculare Aufgaben zum ersten Messzeitpunkt in Paralleltestversion A und die Infit und Outfit-Koeffizienten zu den Messzeitpunkten eins und fünf

Dimension curriculare Aufgaben							
MZP 1				MZP 1		MZP 5	
Item	**Subtest**	**Item-parameter**	**Punkt-biseriale Korrelation**	**OUTFIT MSQ**	**INFIT MSQ**	**OUTFIT MSQ**	**INFIT MSQ**
21	Verdoppeln	-1.45	.46	**0.76***	0.90	**0.54***	**0.61***
22	Verdoppeln	-1.25	.49	**0.76***	**0.89***	0.78	**0.79***
23	Verdoppeln	-0.10	.56	0.84*	0.88*	0.82	0.87*
24	Verdoppeln	0.25	.58	0.83*	0.87*	0.84*	0.89*
25	Verdoppeln	0.81	.51	0.92	0.94	0.99	1.02
26	Verdoppeln	0.82	.53	0.86	0.93	0.84*	0.86*
27	Halbieren	-1.90	.39	0.78	0.91	**0.44***	**0.70***
28	Halbieren	0.62	.53	0.89	0.93	0.97	0.99
29	Halbieren	-0.27	.49	0.88	0.96	0.90	0.89*
30	Halbieren	0.52	.38	1.11	1.07	1.14	1.09
31	Halbieren	1.27	.36	1.16	1.03	1.05	0.99
32	Halbieren	1.66	.32	1.19	1.00	1.03	0.90
33	Ergänzen	-1.50	.35	0.91	0.99	1.28	1.32*
34	Ergänzen	0.74	.30	1.18*	1.14*	0.98	0.97
35	Ergänzen	1.50	.36	1.07	1.03	1.25*	1.05
36	Ergänzen	1.33	.30	1.17	1.07	0.90	0.88*
37	Addition	-1.86	.28	1.10	0.96	**2.02***	**1.51***
38	Addition	-1.91	.33	0.93	0.97	1.12	1.35*
39	Addition	-0.48	.54	0.80*	0.89*	**0.72***	0.83*
40	Addition	1.02	.43	1.02	1.01	0.89	0.95
41	Subtraktion	-1.36	.37	0.93	0.97	0.99	1.12
42	Subtraktion	-1.16	.42	0.86	0.96	0.96	1.07
43	Subtraktion	-0.52	.34	1.08	1.09*	0.94	0.99
44	Subtraktion	0.67	.21	**1.34***	**1.22***	**1.37***	**1.23***
45	Multiplikation	-0.29	.44	0.96	1.01	0.96	1.01
46	Multiplikation	0.61	.46	0.98	0.99	0.88	0.92
47	Multiplikation	0.58	.19	**1.32***	**1.25***	**1.33***	**1.24***
48	Multiplikation	1.66	.11	**1.36***	1.14	1.05	0.97

Anmerkungen. MZP = Messzeitpunkt. *signifikante Abweichung ($p < .05$) vom Mean Square Fit (MSQ); fett markiert sind auffällige signifikante Werte unter 0.8 und über 1.2

wird (Linacre, 2002, 2003). Für High-Stake Tests, welche zur Bewertung von Schülerinnen und Schüler benutzt werden, wird ein strengerer Wertebereich von 0.8 bis 1.2 vorgeschlagen (Wright & Linacre, 1994).

In der Dimension *Vorläuferkompetenzen* ist bezüglich der In- und Outfit-Werte das Item 13 zu MZP1 auffällig, während es zum MZP5 nahe am Erwartungswert von eins liegt. Das Item 13 hat zu MZP1 einen Bodeneffekt, der bei MZP5 nicht mehr vorhanden ist, da alle Personen über eine höhere Personenparameter verglichen mit MZP1 verfügen. Im MZP5 hat das Item 2 einen signifikanten Overfit. Dieses Item ist das leichteste Item der Dimension und wurde nur von 15 Personen zu MZP5 nicht gelöst. Auch die Items 4, 7 und 15 sind leichte Items, welche zu MZP5 nicht mehr ausreichend zwischen den Personen differenzieren.

In der Dimension *curriculare Kompetenzen* weisen die Items 21, 22 und 27 einen Overfit auf und diskriminieren zwischen leistungsstarken und leistungsschwachen Schülern vor allem zu MZP5 zu stark. Die Items 37, 44, 47 und 48 sind schwierige Items, welche auch bereits zu MZP1 nur von wenigen Schülerinnen und Schülern gelöst werden. Die Betrachtung der punktbiserialen Korrelationskoeffizienten legt hier eine geringe Diskriminationsfähigkeit zwischen den Personen (Bodeneffekt) insbesondere im unteren Fähigkeitsbereich nahe.

Vergleich der Itemparameter und der Personenparameter

Für die weiteren Analysen der Personenparameter werden die Items, welche auffällige Infit- und Outfit-Koeffizienten aufwiesen, ausgeschlossen. Die Verteilung der Personenparameter im Vergleich zu den Itemparametern kann anhand der Person-Item–Map dargestelltt werden. Um die Lernentwicklung über die Zeit abzubilden, wurden die Itemparameter aus MZP1 auch für die Schätzung der Personenparameter aus MZP5 zu Grunde gelegt. In Abbildung 2 sind die Personenparameter als Histogramme dargestellt, welche den Itemparametern gegenübergestelt werden.

Für die Dimension *Vorläuferkompetenzen* zeigt sich in Abbildung 2, dass das Fähigkeitsniveau für einen großen Teil der Schülerinnen und Schüler durch die Itemparameter abgedeckt ist, der Test für einen großen Teil der Stichprobe also ein gutes „Targeting" aufweist. Ein geringer Teil der Schülerinnen und Schüler weist jedoch bereits hier sehr hohe Kompetenzen auf, welche nicht mehr von entsprechend schwierigen Items angemessen erfasst werden. Vier Monate später (zum MZP5) sind die Leistungen der Schülerinnen und Schüler erwartungsgemäß höher. Die Aufgaben zu den Vorläuferkompetenzen sind für den Großteil der Schülerinnen und Schüler insgesamt eher zu leicht und das Targeting des Tests verschlechtert sich somit. Für die schwächste Gruppe der Schülerinnen und Schüler sind die Aufgaben allerdings auch zu MZP5 angemessenen. Die WLE-Reliabilität beträgt zu MZP1 .79 und zu MZP5 .73.

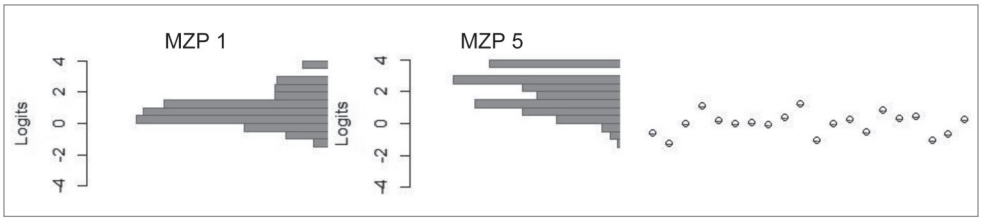

Abbildung 2: Person (MZP 1 und 5) - Item (MZP 1) - Map der Vorläuferkompetenzen zum Vergleich der Kompetenzverteilungen zu den beiden Messzeitpunkten

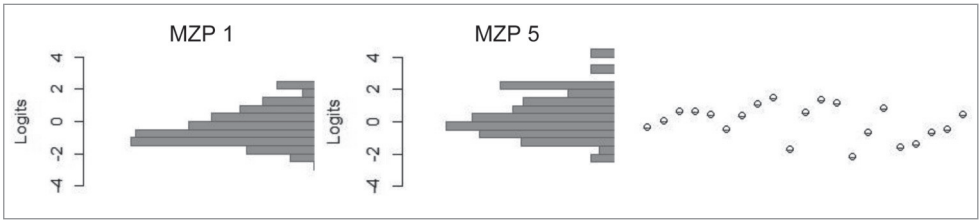

Abbildung 3: Person (MZP1 und 5) - Item (MZP1) - Map der Dimension curriculare Aufgaben zum Vergleich der Kompetenzverteilungen zu den beiden Messzeitpunkten

Für die Dimension *curriculare Aufgaben* liegen die Personenparameter der Schülerinnen und Schüler zu MZP1 auf der Höhe der Itemparameter, was für ein gutes Targeting des Tests zu diesem Zeitpunkt spricht. Zu MZP5 steigt die Kompetenz erkennbar an (vgl. Abbildung 3). Im Vergleich zur Abbildung 2 decken die Items hier insgesamt einen größeren Bereich der Logit-Skala ab. Nur für sehr leistungsstarke Schülerinnen und Schüler finden sich dabei nicht mehr ausreichend schwierige Aufgaben. Die WLE-Reliabilität beträgt bei MZP1 .83 und bei MZP5 .86.

Überprüfung der Testfairness

Analog zu der oben dargestellten Prüfung der Invarianz über die Zeit kann mit dem grafischen Modelltest auch die Subgruppeninvarianz überprüft werden. Dabei werden die Itemparameter jeweils getrennt für den in zwei Teilstichproben aufgeteilten Datensatz berechnet und die so bestimmten Itemparameter jeweils an der X- und Y-Achse in einer Grafik abgetragen (vgl. Abbildung 4). Bei perfekter Modellgeltung sollten sich die abgetragenen Itemparameter entlang der Winkelhalbierenden der Grafik anordnen. Üblicherweise wird zur Überprüfung der Modellgeltung nach diesem Verfahren als Teilungskriterium der Median des Summenwertes der Items des Fragebogens herangezogen (Kubinger, 2005). Nach diesem Verfahren können die Items auch bezüglich anderer Teilungskriterien der Personenstichprobe (z.B. Geschlecht) auf Stichprobeninvarianz ihrer Parameter überprüft werden. Items, die in den verschiedenen Teilstichproben unterschiedliche Schwierigkeiten aufweisen und bei der grafischen Darstellung somit von der Winkelhalbierenden abweichen, widersprechen den Modellannahmen des Raschmodells. In der Abbildung 4 auf der linken Seite wurde der klassische grafische Modelltest mit dem Median als

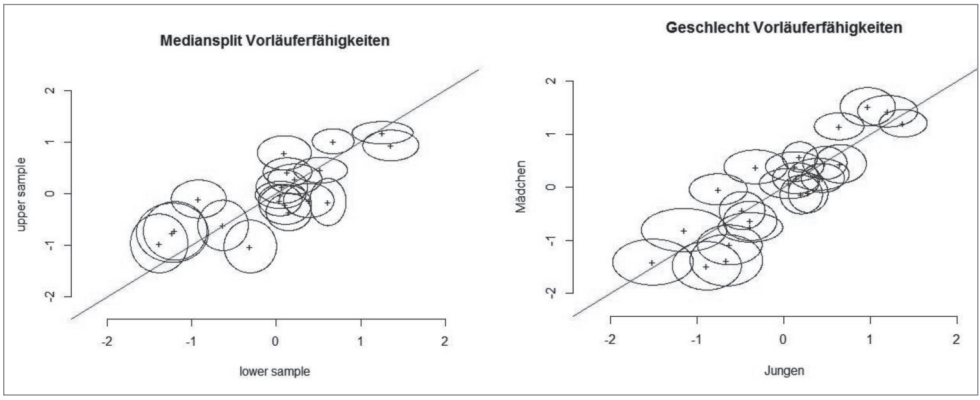

Abbildung 4: Grafischer Modeltest der Vorläuferkompetenzen zum ersten Messzeitpunkt

Teilungskriterium vorgenommen, der zunächst die Raschhomogenität der Items der Skala *Vorläuferkompetenzen* nahelegt. Auf der rechten Seite der Abbildung wurde als Teilungskriterium das Geschlecht herangezogen. Es zeigt sich hier, dass einige der Items geringes Differential Item Functioning (DIF) aufweisen. Für das Teilungskriterium Median (linke Seite) zeigt sich, dass Personen mit niedrigen Personenparametern häufiger als Personen mit hohen Personenparametern die Items 1 und 12 lösen, während die Personen mit hohen Personenparametern die Items 18 und 20 häufiger lösen. In Bezug auf Geschlechtsunterschiede waren die Items 5, 6 und 7 für Jungen leichter, während das Item 50 von Mädchen häufiger gelöst wurde.

In Abbildung 5 sind die beiden grafischen Modelltests zur Überprüfung von DIF der curricularen Aufgaben dargestellt. Im linken Teil erkennt man, dass einige schwierigere Aufgaben (30, 31, 32, 34, 35 und 36) leichter für Personen mit niedrigen Personenparametern zu lösen sind, während einige leichtere Aufgaben (23, 24 und 39) für Personen mit höheren Personenparametern leichter sind. Hier findet sich DIF über einige Aufgaben hinweg. Im Test auf Subgruppeninvarianz nach dem Geschlecht sind zwei Items auffällig: Item 38 wird leichter von Mädchen gelöst, während das Item 34 eher von Jungen richtig beantwortet wird.

Diskussion

Das Ziel der vorgestellten Re-Analyse war es, die Befunde von Salaschek und Souvignier (2014) zu überprüfen und Erkenntnisse über eine mögliche Weiterentwicklung des Instrumentes für einen Einsatz in inklusiven Klassen abzuleiten. Es zeigte sich, dass die von Salaschek und Souvignier (2014) berichtete Zweidimensionalität des Tests durch den Modellvergleich nach der IRT-Skalierung gestützt wird. Die Item Fit-Statistiken sowie die punktbiserialen Korrelationskoeffizienten der beiden eindimensionalen Skalen *Vorläuferkompetenzen* und *curriculare Aufgaben* weisen weitgehend akzeptable Werte auf. Beide Skalen erweisen sich in der Testversion A von MZP1 zu MZP5 als reliabel und sensitiv für Lernzuwächse. Insbesondere die Leistungen schwächerer Schülerinnen und Schüler lassen sich mit dem Instrument reliabel messen.

Vor dem Hintergrund der Zielsetzung, mit einer Reihe von auf Basis der KTT konstruierten Paralleltests Kompetenzen zu messen, deren (nahezu vollständige) Beherrschung erst am Schuljahresende erwartet wird, zeigen die vorliegenden Ergebnisse, dass die Paralleltests dieses Ziel angemessen erfüllen: Sie adressieren die curricular definierten Kompetenzen inhaltlich, während gleichzeitig ein großer Teil der Schülerinnen und Schüler tatsächlich am

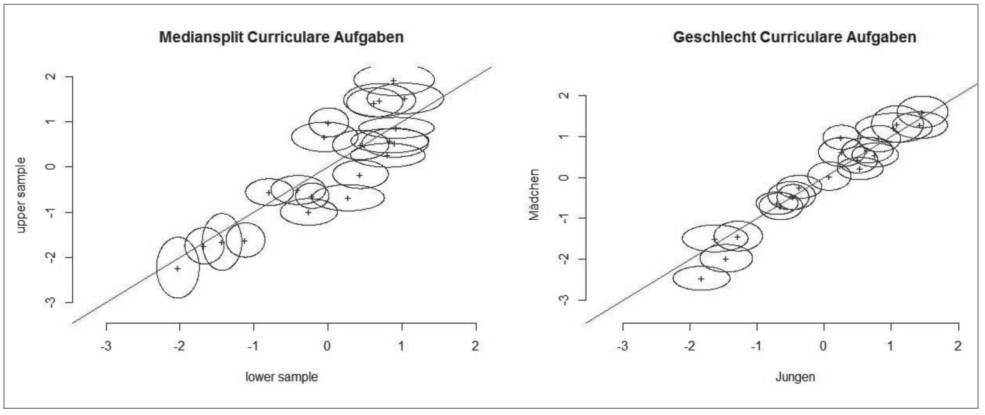

Abbildung 5: Grafischer Modelltest der curricularen Aufgaben

Schuljahresende diese Kompetenzen zu beherrschen scheinen. Allerdings bedeutet dies auch, dass zum Ende des Schuljahres keine Differenzierung zwischen den leistungsstärkeren Schülerinnen und Schülern mehr möglich ist. Die vorliegenden Ergebnisse zum DIF deuten im Allgemeinen auf eine hohe Testfairness hin. Insbesondere die DIF-Befunde für einzelne Items bezüglich des Geschlechts lassen sich inhaltlich nicht erklären. Im Hinblick auf den Befund, dass bei einigen Items leistungsstärkere Schülerinnen und Schüler jedoch größere Schwierigkeiten mit einigen Items und leistungsschwächere weniger Schwierigkeiten haben, als nach ihrer Fähigkeit anzunehmen wäre, liegt eine mögliche Erklärung in den Distraktoren, die eventuell leistungsstarke Schülerinnen und Schüler zu Flüchtigkeitsfehlern verleiten.

Generell erscheinen beide Dimensionen, die durch die Tests erfasst werden, für einen Einsatz in inklusiven Klassen der zweiten Jahrgangsstufe geeignet. Dennoch gibt es einige Möglichkeiten, die Eignung der Tests für einen Einsatz in inklusiven Klassen zu erhöhen. So wäre es wünschenswert, dass sowohl zwischen leistungsstärkeren Schülerinnen und Schülern noch am Schuljahresende ausreichend differenziert werden kann, gleichzeitig aber die Leistungen von Schülerinnen und Schülern mit SPF adäquat abgebildet werden. Items sollten dabei Subgruppeninvarianz aufweisen, so dass beispielsweise leistungsstarke oder -schwache Schülerinnen und Schüler oder auch Mädchen und Jungen nicht unterschiedlich durch einzelne Items benachteiligt werden. Folgende Vorschläge können deshalb aus den vorliegenden Ergebnissen für eine Adaptierung der Testreihe an diese Erfordernisse abgeleitet werden:

- Ergänzung um leichtere Aufgaben, um auch die Leistungen besonders leistungsschwacher Schülerinnen und Schüler schon zu Beginn des Schuljahres abzubilden.
- Ergänzung um schwierigere Aufgaben, die sich allerdings noch innerhalb der curricularen Vorgaben bewegen müssen, um auch am Schuljahresende noch ausreichend zwischen besonders leistungsstarken Schülerinnen und Schülern differenzieren zu können.
- Genauere Betrachtung der Distraktoren als eventuelle Quelle von DIF (insbesondere Vermeidung von Distraktoren, die selbst die Itemschwierigkeit erhöhen; letztere sollte allein auf den Anforderungen der zu leistenden Rechenoperationen basieren und nicht auf einer zu starken oder schwachen Abhebung der Distraktoren untereinander und auch gegenüber der korrekten Lösung).
- Überprüfung der schwierigkeitsgenerierenden Merkmale der Items (vgl. Irvine & Kyllonen, 2002) und darauf basierende regelgeleitete Konstruktion von neuen Items, um möglicherweise die Subtests innerhalb der zwei Dimensionen explizit anhand einheitlicher grundlegender Prinzipien zu verlinken (in diesem Fall wäre auch eine Reduzierung der Subtests denkbar, wenn durch diese dann ebenso adäquat die Kompetenzfacetten erfasst werden können).

Zuletzt soll noch auf zwei Einschränkungen der vorliegenden Studie hingewiesen werden. Ein Nachteil des Studiendesigns besteht darin, dass die Aufgaben bisher nicht zwischen den Paralleltestversionen verlinkt sind. Dadurch war es nicht möglich, die Fairness und die Schwierigkeiten bzw. Itemparameter der Paralleltests untereinander zu vergleichen. Zur weiteren Überprüfung wäre daher in Folgestudien ein Booklet-Design nach einem Incomplete Block Design vorteilhafter. Auf diese Weise werden die Items untereinander verlinkt, wodurch ein Gesamtmodell über alle Messzeitpunkte geschätzt werden könnte. Des Weiteren sollten alle Aufgaben in verschiedenen Gruppen zum ersten Messzeitpunkt eingesetzt werden, damit die Itemparameter getrennt von Lerneffekten bestimmt und gegebenenfalls fixiert werden können (Strathmann & Klauer, 2010; Wilbert & Linnemann, 2011).

Die zweite Einschränkung betrifft die vorliegende Stichprobe: In der vorliegenden Studie wurden keine Schülerinnen und Schüler mit SPF untersucht, sondern es wurden Konsequenzen für Schülerinnen und Schüler mit SPF aus der Betrachtung besonders leistungsschwacher Schülerinnen und Schüler der vorliegenden Stichprobe abgeleitet. Die berichteten Ergebnisse sollten daher in weiteren Studien für Schülerinnen und Schüler mit SPF geprüft werden. Ebenfalls könnten Testungen mit einem breiten Itempool derselben Dimensionen neben der zweiten Jahrgangsstufe auch zu Beginn der dritten Jahrgangstufe der Grundschule durchgeführt werden. Dadurch wäre auch die Entwicklung eines klassenstufenübergreifenden Instruments zur Lernverlaufsdiagnostik für Schülerinnen und Schüler mit (und natürlich auch ohne) SPF möglich.

Literaturverzeichnis

Berch, D. B. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities, 38*, 333-339.

Blumenthal, Y., Kuhlmann, K. & Hartke, B. (2014). Diagnostik und Prävention von Lernschwierigkeiten im Aptitude Treatment Interaction- (ATI) und Response to Intervention- (RTI-)Ansatz. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (Tests und Trends N.F. Band 12) (S. 61–82). Göttingen: Hogrefe.

Bundschuh, K. (2010). *Sonderpädagogische Diagnostik*. Stuttgart: UTB.

Choppin, B. (1968). Item Bank using Sample-free Calibration. *Nature*, *219*, 870-872.

Choppin, B. (1985). A fully conditional estimation procedure for Rasch model parameters. *Evaluation in Education, 9*, 29-42.

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

Förster, N. & Souvignier, E. (2014). Learning progress assessment and goal setting: Effects on reading achievement, reading motivation and reading self-concept. *Learning and Instruction, 32*, 91-100.

Förster, N. & Souvignier, E. (2015). Effects of providing teachers with information about their students' reading progress. *School Psychology Review, 44*, 60-76.

Frey, A. (2012). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 275–293). Berlin, Heidelberg: Springer.

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

Gebhardt, M. Schwab, S., Krammer, M. & Gasteiger-Klicpera, B. (2012). Achievement and integration of students with special needs (SEN) in the fifth grade. *Journal of Special Education and Rehabilitation, 13*, 7-19.

Heimlich, U., Lotter, U. & März, M. (2005). *Diagnose und Förderung im Förderschwerpunkt Lernen. Eine Handreichung für die Praxis.* Donauwörth: Auer.

Heine, J.-H. (2014). *pairwise: Rasch Model Parameters by Pairwise Algorithm* [Computer software]. Munich. Retrieved from http://cran.r-project.org/web/packages/pairwise/index.html (R package version 0.2.5).

Heine, J.-H., Sälzer, C., Borchert, L., Siberns, H. & Mang, J. (2013). Technische Grundlagen des fünften internationalen Vergleichs. In M. Prenzel, C. Sälzer, E. Klieme & O. Köller (Hrsg.), *PISA 2012 - Fortschritte und Herausforderungen in Deutschland*. Münster: Waxmann.

Heine, J. H. & Tarnai, Ch. (2013). *Die Pairwise-Methode zur Parameterschätzung im ordinalen Rasch-Modell*. Vortrag auf der 11. Tagung der Fachgruppe Methoden & Evaluation der DGPs, Klagenfurt, 19.09.2013 - 21.09. 2013.

Heine, J. H. & Tarnai, Ch. (2015). Pairwise Rasch Model Item Parameter Recovery under Sparse Data Condtions. *Psychological Test and Assessment Modeling 57*(1), 3-36.

Huber, C. & Grosche, M. (2012). Das Response-To-Intervention-Modell als Grundlage für einen inklusiven Paradigmenwechsel in der Sonderpädagogik. *Zeitschrift für Heilpädagogik, 63*, 312-322.

Irvine, S. H. & Kyllonen, P. C. (2002). *Item generation for test development*. Mahwah, NJ: Erlbaum.

Kiefer, T., Robitzsch, A. & Wu, M. (2014). *TAM: Test-Analysis Modules*. Retrieved from http://cran.r-project.org/web/packages/TAM/index.html.

Klauer, K. J. (2014). Formative Leistungsdiagnostik. Historischer Hintergrund und Weiterentwicklung zur Lernverlaufsdiagnostik. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (S. 1-19). Weinheim: Hogrefe.

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*, 513-526.

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model – Some critical suggestions on traditional approaches. *International Journal of Testing, 5*, 377–394.

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

Linacre, J.M. (2003) Rasch Power Analysis: Size vs. Significance: Standardized Chi-Square Fit Statistic. *Rasch Measurement Transactions, 17*, 918.

Müller, C. & Hartmann, E. (2009). Lernfortschritte im Unterricht erheben – Möglichkeiten und Grenzen des curriculumbasierten Messens. *Schweizerische Zeitschrift für Heilpädagogik, 15,* 36-42.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rost, J. & Schiefele, U. (2006). *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres*. Münster: Waxmann.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks pædagogiske Institut.

Reschly, A.L., Busch, T.W., Betts, J., Deno, S.L. & Long, J.D. (2009). Curriculum based measurement oral reading as an indicator of reading achievement: A meta analysis of the correlational evidence. *Journal of School Psychology, 47*, 427-269.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion*. Bern: Huber.

Salaschek, M. & Souvignier, E. (2014). Web-based mathematics progress monitoring in second grade. *Journal of Psychoeducational Assessment*, *32*, 710-724.

Schwarz, G. E. (1978). Estimating the dimension of a model, *Annals of Statistics, 6*, 461-464.

Souvignier, E. & Förster, N. (2011). Effekte prozessorientierter Diagnostik auf die Entwicklung der Lesekompetenz leseschwacher Viertklässler. *Empirische Sonderpädagogik, 3,* 243-255.

Souvignier, E., Förster, N. & Salaschek, M. (2014). quop: ein Ansatz internet-basierter Lernverlaufsdiagnostik und Testkonzepte für Mathematik und Lesen. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik*. (Tests und Trends N.F. Band 12) (S. 239-256). 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 School, 42*, 795–819.

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

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

Voß, S. & Hartke, B. (2014). Curriculumbasierte Messverfahren (CBM) als Methode der formativen Leistungsdiagnostik im RTI-Ansatz. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (S. 1-19). Weinheim: Hogrefe.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450.

Walter, J. (2009). *LDL - Lernfortschrittsdiagnostik Lesen. Ein curriculumbasiertes Verfahren*. Weinheim: Hogrefe.

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

Wayman, M., Wallace, T., Wiley, H.I., Ticha, R. & Espin, C.A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education, 41*, 85-120.

Wilbert, J. (2014). Instrumente zur Lernverlaufsmessung Gütekriterien und Auswertungsherausforderungen. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (S. 281–308). Weinheim: Hogrefe.

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

Wocken, H. & Gröhlich, C. (2009). *Kompetenzen von Schülerinnen und Schülern an Hamburger Förderschulen*. In: W. Bos, M. Bonsen & C. Gröhlich (Hrsg.), KESS 7 – Kompetenzen und Einstellungen von SchülerInnenn an Hamburger Schulen zu Beginn der Jahrgangsstufe 7 (S. 133-142). Münster: Waxmann.

Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.

Wright, B. D. & Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.

Markus Gebhardt
TU München
School of Education
Susanne Klatten-Stiftungslehrstuhl für Empirische Bildungsforschung
Arcisstr. 21
80333 München
markus.gebhardt@tum.de