Kocaj, Aleksander; Haag, Nicole; Weirich, Sebastian; Kuhl, Poldi; Pant, Hans Anand; Stanat, Petra

Aspekte der Testgüte bei der Erfassung schulischer Kompetenzen von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf

Moser, Vera [Hrsg.]; Lütje-Klose, Birgit [Hrsg.]: Schulische Inklusion. Weinheim; Basel : Beltz Juventa 2016, S. 212-234. - (Zeitschrift für Pädagogik, Beiheft; 62)

62. Beiheft April 2016

ZEITSCHRIFT FÜR PÄDAGOGIK

Schulische Inklusion

BELTZ JUVENTA

Zeitschrift für Pädagogik · 62. Beiheft

Zeitschrift für Pädagogik · 62. Beiheft

Schulische Inklusion

Herausgegeben von
Vera Moser und Birgit Lütje-Klose

BELTZ JUVENTA

www.beltz.de · www.juventa.de
Herstellung: Lore Amann
Satz: text plus form, Dresden

E-Book

ISSN 0514-2717
Bestell-Nr. 443510

Inhaltsverzeichnis

Aleksander Kocaj/Nicole Haag/Sebastian Weirich/Poldi Kuhl/
Hans Anand Pant/Petra Stanat

Aspekte der Testgüte bei der Erfassung schulischer Kompetenzen von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf

Zusammenfassung: Im vorliegenden Beitrag wurde geprüft, inwieweit die im IQB-Ländervergleich 2011 in der Primarstufe eingesetzten Kompetenztests in Deutsch und Mathematik dazu geeignet sind, auch die Kompetenzen von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf (SPF) adäquat zu erfassen. Dazu wurden der Anteil fehlender Werte, die Passung des Raschmodells, differenzielle Itemfunktionalität und Korrelationen der Testwerte zwischen den Schülergruppen mit SPF in Förder- und Grundschulen und Kindern ohne SPF analysiert. Die eingesetzten Tests scheinen die schulischen Kompetenzen von Kindern mit SPF reliabel und valide zu erfassen. Allerdings zeigten sich Einschränkungen hinsichtlich der Vergleichbarkeit der Messungen für Kinder in Förderschulen in Mathematik.

Schlagworte: Sonderpädagogischer Förderbedarf, Vergleichbarkeit, Schulische Kompetenzen, Primarstufe, Beschulungsart

1. Einleitung

Large-scale assessments (LSAs) im Rahmen des Bildungsmonitorings zielen darauf ab, schulische Kompetenzen von Schülerinnen und Schülern zu messen und Unterschiede in den Erträgen von Bildungssystemen insgesamt sowie für verschiedene Gruppen von Lernenden sichtbar zu machen. Damit sollen die Studien Stärken und Schwächen auf der Systemebene identifizieren und Impulse für eine Verbesserung der Bedingungen schulischen Lernens geben (Chudowsky & Pellegrino, 2003). Diese Konzeptualisierung von LSAs setzt voraus, dass ihre Ergebnisse die jeweilige Zielpopulation und relevante Teilpopulationen adäquat repräsentieren (Olson & Goldstein, 1997).

Eine Schülergruppe, die in den letzten Jahren verstärkt in den Fokus der empirischen Bildungsforschung gerückt ist, ist die der Schülerinnen und Schüler, bei denen ein sonderpädagogischer Förderbedarf (SPF) diagnostiziert wurde. Diese Gruppe wird im Rahmen von nationalen und internationalen LSAs (z. B. Ländervergleich des IQB, PISA, IGLU) soweit wie möglich als Teil der Zielpopulation definiert[1] und in die Stichproben-

1 Vorab aus der Zielpopulation ausgeschlossen werden Schülerinnen und Schüler, die aufgrund sehr geringer Sprachkenntnisse (Beschulungsdauer in der Testsprache von weniger als einem Jahr) oder einer Beeinträchtigung nicht in der Lage wären, die Tests zu bearbeiten.

ziehung eingeschlossen. Damit soll gewährleistet werden, dass die Schülerschaft, auf die sich das Monitoring bezieht, möglichst breit erfasst wird.

Die Einbeziehung von Schülerinnen und Schülern mit SPF im Bildungsmonitoring ist aber vor allem auch aufgrund des steigenden Anteils von Schülerinnen und Schülern mit SPF an Regelschulen des allgemeinbildenden Schulsystems relevant (Autorengruppe Bildungsberichterstattung, 2014). Die UN-Behindertenrechtskonvention (UN-BRK) weist darauf hin, dass empirische Daten bereitgestellt werden müssen, um die Entwicklung eines inklusiven Bildungssystems evaluieren zu können und der Bildungspolitik Anhaltspunkte für eine effektive Steuerung der Umsetzung dieser Reform zu liefern (Beauftragter der Bundesregierung für die Belange behinderter Menschen, 2010, Artikel 31). Zur Umsetzung der auf Inklusion bezogenen rechtlichen Vorgaben ist es also wichtig, Schülerinnen und Schüler mit SPF in Schulleistungsstudien einzubeziehen (Thurlow, 2002). LSAs können als wichtiges Instrument zur Evaluation inklusiver Bemühungen im deutschen Bildungssystem genutzt werden (Wrase, 2015), um die schulische Entwicklung von Kindern mit SPF zu verfolgen (Abedi et al., 2011).

Aufgrund dieser Zielperspektiven wird zunehmend versucht, die Einbeziehung von Kindern und Jugendlichen mit SPF in nationalen und internationalen Schulleistungsstudien zu optimieren (Thurlow, 2010). Im Fokus nationaler Schulleistungsstudien stehen dabei vor allem Schülerinnen und Schüler mit den Förderschwerpunkten *Lernen, Sprache* sowie *emotionale und soziale Entwicklung.* Dabei handelt es sich – neben *geistiger Entwicklung* – um die drei größten Förderschwerpunkte in Deutschland (Autorengruppe Bildungsberichterstattung, 2014), wobei die Differenzierung zwischen den Förderschwerpunkten aufgrund von substanziellen Überschneidungen in Einzelfällen schwierig ist (Opp, Budnik & Fingerle, 2008; Ricking, 2005). Auch ist die Abgrenzung dieser drei Schülergruppen mit SPF von Schülerinnen und Schülern ohne SPF nur graduell möglich und scheint zum Teil durch soziale Bezugsnormen beeinflusst zu werden (Bos, Müller & Stubbe, 2010; Kornmann, 2006). Für diese Schülergruppen wird daher zunehmend diskutiert, auf eine separate Beschulung in Förderschulen vollständig zu verzichten und sie stattdessen in Regelschulen des allgemeinbildenden Schulsystems zu unterrichten (Autorengruppe Bildungsberichterstattung, 2014).

Die Kompetenztestung von Schülerinnen und Schülern mit SPF in LSAs wird kontrovers diskutiert (Smith & Douglas, 2014; Thurlow, 2002). Die primäre Funktion von LSAs besteht in der Evaluation von schulischen Outcomes auf der Systemebene. Impulse für die Unterrichtsentwicklung und die konkrete Ausgestaltung von Inklusion sind somit nicht Ziel von LSAs. Allerdings befürchten Kritiker, dass die für LSAs konzipierten Tests auch für diese Zwecke verwendet werden könnten (Smith & Douglas, 2014). Dies würde zu einer zu starken Betonung schulischer Leistungen bei der Evaluation inklusiver Bemühungen führen (Smith & Douglas, 2014). Darüber hinaus ermöglicht die überwiegend querschnittliche Ausrichtung von LSAs nicht, Leistungsentwicklungen von Schülerinnen und Schülern mit SPF abzubilden. Im Rahmen eines inklusiven Ansatzes wird argumentiert, dass weitere schulische Outcomes (z. B. soziale Kompetenzen, Motivation, Wohlbefinden) und Entwicklungsverläufe von Schülerinnen und Schülern mit SPF betrachtet werden sollten (Smith & Douglas, 2014). Auf der anderen

Seite stellen LSAs eine wichtige Grundlage dar, um den Bildungserfolg der Schülergruppe mit SPF im Hinblick auf die weitere schulische und berufliche Ausbildung zu evaluieren. Zudem könnten hohe Leistungserwartungen an alle Schülerinnen und Schüler die Erwartungshaltung und Verantwortungsbereitschaft der Lehrkräfte erhöhen, wodurch auch die Schülergruppe mit SPF in ihrer Leistungsentwicklung profitieren könnte (Thurlow, 2002). Schließlich besteht ein hoher gesellschaftlicher Konsens darüber, dass Schülerinnen und Schülern mit SPF Zugang zu den gleichen schulischen Erfahrungen ermöglicht werden sollte, die auch Kinder ohne SPF erfahren (Pitoniak & Royer, 2001). LSAs ersetzen somit nicht die Erfassung individueller Lernfortschritte in akademischen, motivationalen und sozialen Bereichen, sie können aber ein nützliches Instrument zur Evaluation des schulischen Abschneidens von Kindern mit SPF darstellen.

Um aus dem Abschneiden von Schülerinnen und Schülern mit SPF in LSAs Schlussfolgerungen für inklusive Bemühungen im deutschen Bildungssystem ableiten zu können, muss zunächst geprüft werden, inwieweit die derzeit eingesetzten Kompetenztests für diese Schülergruppe geeignet sind. Die Tests wurden ursprünglich für Schülerinnen und Schüler ohne SPF entwickelt und erprobt, sie werden jedoch auch zur Erfassung der Kompetenzen von Schülerinnen und Schülern mit einem Förderbedarf in den Bereichen *Lernen, Sprache* oder *emotionale und soziale Entwicklung* eingesetzt. Im Folgenden soll der Frage nachgegangen werden, inwieweit die Kompetenztests dafür geeignet sind, wobei die Prüfung sowohl für Kinder mit SPF an Förderschulen als auch für Kinder mit SPF an Grundschulen vorgenommen wird.

2. Methodische Überlegungen

Die zentrale Herausforderung bei der gemeinsamen Testung von Schülerinnen und Schülern mit und ohne SPF besteht darin, die schulischen Leistungen der gesamten Schülerschaft auf einer gemeinsamen Skala abzubilden und gleichzeitig für jede Schülergruppe eine reliable und valide Testung zu gewährleisten (Heydrich, Weinert, Nusser, Artelt & Carstensen, 2013). Damit verbunden ist die Frage der Äquivalenz der Testergebnisse von Heranwachsenden mit SPF und Heranwachsenden ohne SPF (Lane & Leventhal, 2015). Eine eingeschränkte Vergleichbarkeit der Testergebnisse könnte dazu führen, dass Unterschiede zwischen diesen beiden Schülergruppen im interessierenden Kompetenzkonstrukt (z. B. mathematische Kompetenz) durch die Testwertunterschiede nicht oder nur verzerrt abgebildet werden.

Um die Testgüte bzw. Validität schulischer Leistungstests zu beurteilen, muss geprüft werden, ob die Testwerte die zugrundeliegenden Fähigkeiten adäquat erfassen und ob die Interpretation der Testwerte für verschiedene Schülergruppen vergleichbar ist (Messick, 1995). In der Literatur werden verschiedene Aspekte der Validität diskutiert (Messick, 1995). In der vorliegenden Untersuchung wird die *Generalisierbarkeit* der Interpretation von Testwerten aus Schulleistungserhebungen auf die Teilpopulation der Schülerinnen und Schüler mit SPF in den Bereichen *Lernen, Sprache* sowie *emotionale und soziale Entwicklung* in den Blick genommen. Es wird geprüft, ob die

Gütekriterien der Kompetenztests für Kinder mit SPF in Förder- und Grundschulen in vergleichbarer Weise wie für Kinder ohne SPF erfüllt sind und inwieweit sich die Interpretation der Testwerte bei Kindern ohne SPF auf die Schülergruppe mit SPF übertragen lässt. Zur Beurteilung der Generalisierbarkeit werden strukturelle und externe Aspekte der Validität betrachtet. Die Vergleichbarkeit der Aussagekraft von Testwerten für verschiedene Schülergruppen als ein Merkmal der Testgüte bzw. Validität wird auch unter dem Stichwort der „Testfairness" diskutiert (Cole & Zieky, 2001). Bleiben Aspekte der Testfairness ungeprüft, könnten beobachtete Gruppenunterschiede nicht allein auf Unterschiede zwischen Schülergruppen im zugrundeliegenden Konstrukt, sondern auch auf Merkmale der Testinstrumente zurückzuführen sein (Cole & Zieky, 2001). Die Validität der Ergebnisse bzw. deren Interpretation wäre also eingeschränkt.

Die angemessene Interpretation von Gruppenunterschieden in Testergebnissen setzt voraus, dass strukturelle Aspekte der Testverfahren zwischen Schülergruppen vergleichbar sind (Messick, 1995). Insbesondere sollten die interne Struktur der Leistungstests und die psychometrischen Eigenschaften der Testitems zwischen den Schülergruppen hinreichend ähnlich sein. Strukturelle Aspekte und psychometrische Eigenschaften der im Folgenden analysierten Kompetenztests wurden für Schülerinnen und Schüler ohne SPF bereits bestimmt und validiert (Böhme & Robitzsch, 2009; Winkelmann & Robitzsch, 2009). Diese Kennwerte dienen in der vorliegenden Untersuchung als Referenz zur Überprüfung der Testgüte für Kinder mit SPF in Förder- und Grundschulen. Anhand der bereits vorliegenden Auswertungsmodelle wird geprüft, ob die Kompetenztests in vergleichbarer Weise geeignet sind, die Kompetenzen von Schülerinnen und Schülern mit SPF an Förder- und Grundschulen zu erfassen wie für Schülerinnen und Schüler ohne SPF. Nur wenn gesichert ist, dass die Kompetenztests bei Kindern mit und bei Kindern ohne SPF eine ähnliche Struktur aufweisen, kann vorausgesetzt werden, dass die in LSAs eingesetzten Testverfahren die schulischen Kompetenzen beider Schülergruppen gleichermaßen reliabel und valide messen.

Neben der internen Struktur des Tests sollten die Zusammenhangsmuster der Kompetenztestwerte mit relevanten anderen Konstrukten für die betrachteten Schülergruppen ähnlich ausfallen. Die theoriegeleitete Überprüfung von Zusammenhängen der Testergebnisse mit externen Kriterien, wie z. B. den Ergebnissen anderer Leistungstests, ist ein wichtiger Aspekt der Validierung eines Testverfahrens (Messick, 1995). Zeigen sich ähnliche Zusammenhänge für die untersuchten Schülergruppen, können deren Leistungen auf einer gemeinsamen Skala abgebildet und sinnvoll vergleichend interpretiert werden.

2.1 Methodische Herausforderungen bei der Interpretation der Testwerte von Schülerinnen und Schülern mit SPF in LSAs

Die Übertragbarkeit der Interpretation von Testwerten auf Schülerinnen und Schüler mit SPF kann durch mehrere Aspekte eingeschränkt sein. Ein Aspekt ist die Reliabilität der Leistungstests, die bei dieser Schülergruppe möglicherweise geringer ist. Allge-

mein fällt die Messgenauigkeit von Leistungstests, die auf Modellen der *Item Response Theory* basieren, im mittleren Bereich der Fähigkeitsverteilung am höchsten aus und nimmt zu den Rändern der Fähigkeitsverteilung hin ab (Embretson & Reise, 2000). Für Schülerinnen und Schüler mit SPF ist im Vergleich zu Schülerinnen und Schülern ohne SPF eine deutlich geringere mittlere Fähigkeit (Abedi et al., 2011) sowie eine eingeschränkte Varianz in der Testleistung (Lane & Leventhal, 2015) zu erwarten. Dies kann die Messgenauigkeit der Leistungstests für Schülerinnen und Schüler mit SPF reduzieren.

Als weiterer Aspekt sind mögliche Einschränkungen der Testfairness in Betracht zu ziehen, die sich auf die Interpretierbarkeit der Testleistungen von Schülerinnen und Schülern mit SPF auswirken können. Die eingesetzten Testverfahren sollten es allen Schülerinnen und Schülern ermöglichen, ihre schulischen Kompetenzen zu zeigen. Treten unter Berücksichtigung von Fähigkeitsunterschieden noch differenzielle Schwierigkeiten der Items zwischen Schülergruppen auf, kann die vergleichende Interpretierbarkeit der Testwerte eingeschränkt sein (Abedi et al., 2011). Analysen zur sogenannten „differenziellen Itemfunktionalität" (DIF) (Differential Item Functioning; Holland & Thayer, 1988, Übers. d. Verf.) können Hinweise darauf geben, inwieweit die Lösungswahrscheinlichkeit der einzelnen Items nicht nur durch die schulischen Kompetenzen beeinflusst wird, die erfasst werden sollen, sondern auch durch konstrukt-ferne Merkmale. DIF liegt vor, wenn Schülerinnen und Schüler aus verschiedenen Gruppen trotz gleichem Fähigkeitsniveau in der betrachteten Kompetenz ein spezifisches Item mit unterschiedlich hoher Wahrscheinlichkeit lösen. Bei den konstrukt-fernen Merkmalen, die zu systematischen Unterschieden in der Testbearbeitung und somit zu DIF führen können, kann es sich z. B. um sprachliche Anforderungen, Kontextualisierungen oder Antwortformate der Aufgaben handeln (Abedi et al., 2011; Heydrich et al., 2013).

2.2 Anpassung von Testverfahren für Schülerinnen und Schüler mit SPF in LSAs

Um die Eignung von in LSAs eingesetzten Leistungstests für Schülerinnen und Schüler mit SPF zu verbessern und den Einfluss konstrukt-irrelevanter Merkmale auf die Leistung bei einzelnen Items bzw. im Test insgesamt zu verringern, können sogenannte Akkommodationen vorgenommen werden (Pitoniak & Royer, 2001). Akkommodationen sind Anpassungen in der Testgestaltung, die dazu dienen sollen, diejenige konstrukt-irrelevante Varianz in den Testwerten zu reduzieren, die mit den Beeinträchtigungen der Kinder mit SPF zusammenhängt, ohne die Definition und Operationalisierung des zugrundeliegenden Konstrukts zu verändern (Elliott, Beddow, Kurz & Kettler, 2011).

Eine häufig eingesetzte Akkommodation betrifft die Testlänge. Die Teilnahme an Leistungstests in LSAs erfordert ein ausreichendes Maß an Konzentration und Aufmerksamkeit über einen längeren Zeitraum. Durch kürzere Tests für Schülerinnen und Schüler mit SPF können mögliche Nachteile dieser Schülergruppe in der Aufmerksamkeit und Konzentrationsfähigkeit berücksichtigt werden (Heydrich et al., 2013). Als In-

dikator dafür, ob die Bearbeitungszeit für Kinder mit SPF angemessen ist, kann die Anzahl fehlender Werte herangezogen werden. Insbesondere die Analyse fehlender Werte, die am Ende des Tests auftreten *(Missing not Reached),* kann Hinweise darauf geben, ob Schülerinnen und Schüler mit SPF durch die Länge des Tests in besonderem Maße überfordert waren.

Eine weitere Akkommodation, die in LSAs zum Einsatz kommt, besteht darin, besonders schwierige Items bzw. Testblöcke aus den Testheften für Schülerinnen und Schülern mit SPF zu entfernen (Südkamp, Pohl, Hardt, Jordan & Duchhardt, 2015). Durch den Einsatz leichterer Aufgaben kann eine höhere Passung der Itemschwierigkeiten und Personenfähigkeiten erreicht werden, wodurch sich die Reliabilität der Kompetenztests für Schülerinnen und Schüler mit SPF erhöht (Lane & Leventhal, 2015).

2.3 Forschungsstand zur Interpretation der Testwerte von Schülerinnen und Schülern mit SPF in LSAs

Im Rahmen nationaler Erhebungen wurde nur selten überprüft, ob die Abbildung der Testwerte auf einer gemeinsamen Skala gerechtfertigt ist, da die Testscores von Heranwachsenden mit SPF und Heranwachsenden ohne SPF in vergleichbarer Weise interpretiert werden können. Eine Ausnahme ist die im Rahmen des Nationalen Bildungspanels (NEPS; Südkamp, Pohl, Hardt et al., 2015; Südkamp, Pohl & Weinert, 2015) durchgeführte Zusatzerhebung an Förderschulen.

Südkamp, Pohl und Weinert (2015) untersuchten, inwieweit Leseverständnistests in NEPS für Schülerinnen und Schüler an Förderschulen mit dem Förderschwerpunkt *Lernen* eine reliable und vergleichbare Messung ermöglichen. Als Vergleichsgruppe wurden Schülerinnen und Schüler an Hauptschulen betrachtet. Beide Schülergruppen besuchten die fünfte Klasse. Außerdem wurde geprüft, inwieweit Testakkommodationen zu einer verbesserten Messung der Schülergruppe mit SPF führen. Die Testgüte wurde anhand des Anteils fehlender Werte, des Itemfits, der Reliabilität und anhand von DIF-Analysen evaluiert. Im Standardtest ohne Akkommodationen zeigten sich höhere Anteile fehlender Werte für Schülerinnen und Schüler mit dem Förderschwerpunkt *Lernen* als bei Schülerinnen und Schülern ohne SPF. Im Hinblick auf den Itemfit wies ein bedeutsamer Anteil der Items für Schülerinnen und Schüler mit dem Förderschwerpunkt *Lernen* eine schlechte Passung auf. Außerdem war die Reliabilität des Leseverständnistests deutlich geringer als für Hauptschülerinnen und -schüler. Im Hinblick auf die Testfairness zeigte sich bei einem bedeutsamen Anteil der eingesetzten Aufgaben ein starker DIF. Insgesamt war die Vergleichbarkeit der Testwerte zwischen den betrachteten Schülergruppen somit stark eingeschränkt. Der Einsatz von Akkommodationen (reduzierte Testzeit bzw. einfachere Aufgaben) führte zu einer verbesserten Abbildung der schulischen Kompetenzen (s. a. Südkamp, Pohl, Hardt et al., 2015).

Die beschriebene Studie zur Frage der äquivalenten Interpretierbarkeit der Testwerte bezog sich auf Schülerinnen und Schüler mit SPF in Förderschulen. Im Rahmen von LSA werden jedoch auch die Leistungen von Schülerinnen und Schüler mit SPF

in sonstigen allgemeinbildenden Schulen erfasst. Zudem besteht erhebliches wissenschaftliches und bildungspolitisches Interesse an der Frage, inwieweit die Leistungsentwicklung von Schülerinnen und Schülern mit SPF durch die Art der Beschulung beeinflusst wird (Kocaj, Kuhl, Kroth, Pant & Stanat, 2014). Daher ist es wichtig zu prüfen, ob die eingesetzten Testverfahren Fähigkeitsunterschiede bei Kindern mit SPF in beiden Beschulungsarten gleichermaßen valide und reliabel abbilden können. Außerdem untersuchte die Mehrzahl der Studien Schülerinnen und Schüler in der Sekundarstufe, wohingegen sich die vorliegende Studie auf Schülerinnen und Schüler am Ende der Primarstufe bezieht.

3. Fragestellungen

In der vorliegenden Studie soll geprüft werden, inwieweit die im Rahmen des nationalen Bildungsmonitoring in der Primarstufe eingesetzten Schulleistungstests in den Fächern Deutsch und Mathematik dazu geeignet sind, die Kompetenzen von Schülerinnen und Schülern mit SPF in Förder- und Grundschulen zu erfassen (vgl. IQB-Ländervergleich 2011, Stanat, Pant, Böhme & Richter, 2012). In der 2011 durchgeführten Studie unterschieden sich die Testbedingungen zwischen den beiden Schularten: In Förderschulen kamen Testhefte mit einer reduzierten Anzahl von Aufgabenblöcken und einer entsprechend verkürzten Bearbeitungszeit zum Einsatz. Zudem enthielten die Testhefte eine Auswahl leichterer Aufgaben (s. Abschnitt 4.2). Auf der Grundlage dieser Daten werden vier Forschungsfragen untersucht.

1) *Anteil fehlender Werte: Ist die Bearbeitungszeit für Kinder mit SPF angemessen?* Anhand eines Vergleichs des Anteils fehlender Werte wird zunächst untersucht, ob die für die Kompetenztests zur Verfügung stehende Bearbeitungszeit für die drei Schülergruppen gleichermaßen ausreichend ist. Es wird angenommen, dass der Anteil fehlender Werte für Kinder mit SPF deutlich höher ist als für Kinder ohne SPF. Aufgrund der Anpassungen der in Förderschulen eingesetzten Tests sollte jedoch der Anteil nicht bearbeiteter Aufgaben bei Kindern mit SPF in Förderschulen geringer sein als bei Kindern mit SPF in Grundschulen.
2) *Passung des Raschmodells: Treffen die Annahmen eines eindimensionalen Konstrukts für Kinder mit SPF in vergleichbarer Weise zu wie für Kinder ohne SPF?* Im Hinblick auf die strukturelle Validität der Kompetenztests wird für die einzelnen Kompetenzdomänen geprüft, inwiefern sich die Annahme eines eindimensionalen Konstrukts in allen drei Schülergruppen bestätigen lässt. Wir nehmen an, dass sich auch die schulischen Kompetenzen der Kinder mit SPF in Förder- und Grundschulen mit dem Raschmodell in vergleichbarer Weise abbilden lassen.
3) *Differenzielle Itemfunktionalität: Sind die eingesetzten Aufgaben für Kinder mit SPF differenziell schwieriger als für Kinder ohne SPF?* Um Aspekte der Testfairness zu untersuchen, wird in einem nächsten Schritt geprüft, ob einzelne Items in den drei Gruppen nach Berücksichtigung von Fähigkeitsunterschieden differenzielle Schwie-

rigkeiten aufweisen. Wir gehen davon aus, dass die Aufgaben für Kinder mit SPF – insbesondere in Förderschulen – zwar deutlich schwieriger sind als für Kinder ohne SPF, dass nach Kontrolle von Fähigkeitsunterschieden zwischen den Schülergruppen darüber hinaus aber keine bedeutsamen Differenzen in den Itemschwierigkeiten bestehen.

4) *Vergleich der Korrelationsmuster zwischen den Schülergruppen: Lässt sich die Interpretation der Testwerte im Hinblick auf Zusammenhänge mit externen Kriterien bei Kindern ohne SPF auf die Schülergruppe mit einem SPF übertragen?* Abschließend wird ein Aspekt der externen Validität der Testverfahren geprüft, indem korrelative Zusammenhänge der Testwerte untereinander und mit den kognitiven Grundfähigkeiten zwischen den drei Schülergruppen verglichen werden. Für diese Variablen wird eine empirisch bewährte Korrelationsstruktur angenommen, die sich zwischen den drei Schülergruppen nicht bedeutsam unterscheiden sollte. So sollten die Kompetenzen im Lesen und Zuhören höher miteinander korrelieren als mit den mathematischen Kompetenzen.

4. Methoden

4.1 Stichprobe

Die in den folgenden Analysen verwendeten Daten wurden im Rahmen der 2011 durchgeführten Ländervergleichsstudie des IQB erhoben. Diese Querschnittstudie diente der Überprüfung des Erreichens der Bildungsstandards am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Die gezogene Stichprobe der Viertklässlerinnen und Viertklässler (N = 27081) war bundesweit und für jedes einzelne Bundesland repräsentativ.

Insgesamt umfasste die Stichprobe auch 1195 Schülerinnen und Schüler mit einem diagnostizierten SPF im Bereich *Lernen, Sprache* oder *emotionale und soziale Entwicklung* (durchschnittliches Alter: 10.90 Jahre; SD = 0.64; 66.3 % Jungen). Die Mehrheit dieser Gruppe (n = 762; 63.8 %) besuchte eine Grundschule, während 433 Kinder mit SPF in Förderschulen unterrichtet wurden (36.2 %). In die hier berichteten Analysen wurden nur Schülerinnen und Schüler einbezogen, die an den Kompetenztestungen teilgenommen hatten. Im Fach Deutsch bearbeiteten N = 1071 und im Fach Mathematik N = 1053 Schülerinnen und Schüler mit SPF die Kompetenztests (s. für eine differenzierte Darstellung nach Schulart und Förderschwerpunkt Tab. 1). Die Schülergruppe ohne SPF umfasste 25706 Schülerinnen und Schüler (durchschnittliches Alter: 10.43 Jahre; SD = 0.49; 50 % Jungen).

Schülergruppe	Fach	$N_{insgesamt}$	$N_{Grundschule}$	$N_{Förderschule}$
Schülerinnen und Schüler ohne SPF[1]	Deutsch	25 283	25 283	–
	Mathematik	25 266	25 266	–
Schülerinnen und Schüler mit SPF insgesamt	Deutsch	1 071	658	413
	Mathematik	1 053	659	394
Schülerinnen und Schüler mit FS[2] *Lernen*	Deutsch	536	278	258
	Mathematik	523	282	241
Schülerinnen und Schüler mit FS *Sprache*	Deutsch	301	166	135
	Mathematik	301	167	134
Schülerinnen und Schüler mit FS *emotionale und soziale Entwicklung*	Deutsch	234	214	20
	Mathematik	229	210	19

[1] sonderpädagogischer Förderbedarf, [2] Förderschwerpunkt

Tab. 1: Stichprobenumfänge in den drei untersuchten Schülergruppen

4.2 Instrumente

Schulische Kompetenzen. Die schulischen Kompetenzen der Kinder wurden mit standardisierten Leistungstests in den Fächern Deutsch und Mathematik erfasst, wobei ein Multi-Matrix-Design zur Anwendung kam, d. h. jeder Schüler bekam ein Testheft mit einer Teilmenge aller Aufgaben. Die Testhefte sind so erstellt, dass innerhalb einer Klasse das Gesamt der Testaufgaben durch eine hinreichend große Teilmenge von Schülerinnen und Schülern bearbeitet wird, um gesicherte Aussagen über Leistungen auf Individual- und Klassenebene zu ermöglichen. Im Fach Deutsch wurden Kompetenzen in den Bereichen Lesen und Zuhören getestet. Der Kompetenztest im Fach Mathematik erfasste fünf inhaltsbezogene Teilkompetenzen, die zu einem Globalwert mathematischer Kompetenz zusammengefasst wurden.

Bei der Testdurchführung in Förderschulen kamen verschiedene Akkommodationen in den Testbedingungen zum Einsatz. Die Testhefte unterschieden sich in der Bearbeitungszeit sowie in der Anzahl und dem Schwierigkeitsgrad der eingesetzten Aufgaben. Die Testzeit betrug in Grundschulen 80, in Förderschulen 40 Minuten pro Fach. Damit einhergehend bearbeiteten Schülerinnen und Schüler in Förderschulen eine entsprechend geringere Anzahl von Aufgaben (s. Tab. 3). Die durchschnittliche Zeit, die den Kindern pro Item zur Verfügung stand, war in Förder- und Grundschulen vergleichbar. Außerdem wurde in Förderschulen eine Auswahl der leichteren Aufgaben aus dem gemeinsamen Aufgabenpool verwendet. Es wurden also keine neuen Testaufgaben für Förderschulen entwickelt, sondern Aufgaben eingesetzt, die in Voruntersuchungen mit Kindern ohne SPF hohe Lösungswahrscheinlichkeiten aufwiesen. Somit waren die Testbedingungen für Kinder mit SPF je nach Beschulungsart unterschiedlich. Schüle-

rinnen und Schüler mit SPF in Grundschulen erhielten die gleichen Testhefte wie Kinder ohne SPF und wurden unter denselben Bedingungen getestet, erfuhren also keine Akkommodationen.

Kognitive Grundfähigkeiten. Die kognitiven Grundfähigkeiten wurden mit zwei Untertests des Kognitiven Fähigkeitstests (KFT 4-12+R, Heller & Perleth, 2000) erhoben. Dabei diente die Subskala *Wortschatz* (Cronbachs $\alpha = .72$ für Kinder ohne SPF, $\alpha = .79$ für Kinder mit SPF in Grundschulen, $\alpha = .78$ für Kinder mit SPF in Förderschulen) als Indikator für die verbalen kognitiven Fähigkeiten und die Subskala *Figurenanalogien* ($\alpha = .92$ für Kinder ohne SPF, $\alpha = .92$ für Kinder mit SPF in Grundschulen, $\alpha = .88$ für Kinder mit SPF in Förderschulen) als Indikator für die nonverbalen kognitiven Fähigkeiten der Schülerinnen und Schüler.

4.3 Datenanalyse

Im Folgenden werden die Analysemethoden für jede der vier Forschungsfragen erläutert.

(1) *Anteil fehlender Werte.* Zur Analyse fehlender Werte in den Kompetenztests wurden drei Kategorien unterschieden: *Missing by Intention, Missing Invalid Response* und *Missing not Reached.* Wenn eine Schülerin bzw. ein Schüler bei einem Item keine Antwort gegeben hatte, wurde die Kategorie *Missing by Intention* vergeben. Lag eine Antwort vor, war aber nicht interpretierbar, wurde diese Antwort als *Missing Invalid Response* kodiert. Die Kategorie *Missing not Reached* wurde für fehlende Itemantworten am Ende jedes Testblocks vergeben, wenn die letzten Items des Blocks durchgängig nicht beantwortet wurden und daher vermutet werden konnte, dass das Kind keine Möglichkeit hatte, die letzten Items zu bearbeiten.

(2) *Passung des Raschmodells.* Zur separaten Analyse der Itemantworten in den drei betrachteten Schülergruppen wurde ein eindimensionales Modell der *Item Response Theory* (IRT) (Raschmodell, vgl. Embretson & Reise, 2000) verwendet. Dabei wurden in einem ersten Schritt unter Verwendung der *Marginal Maximum Likelihood*-Schätzmethode die Itemschwierigkeiten bestimmt. In einem zweiten Schritt wurde unter Fixierung der Itemparameter auf die im ersten Schritt geschätzten Werte die Verteilung der latenten Personenfähigkeiten geschätzt. Fehlende Werte *(Missing by Intention)* in den Kompetenztests wurden sowohl bei der Schätzung der Itemparameter als auch bei der Schätzung der Personenfähigkeit als falsche Antworten kodiert.

Die Testitems wurden für jede betrachtete Schülergruppe separat skaliert; anschließend wurde die Passung des Modells für jede Schülergruppe geprüft. Zur Beurteilung der Passung wurde der standardisierte *Weighted Mean Square* (WMNSQ, Wright & Masters, 1982) für jedes Item herangezogen. Dieser Kennwert beschreibt die Abweichung der beobachteten Lösungswahrscheinlichkeit eines Items von der durch das Modell implizierten Wahrscheinlichkeit für ein bestimmtes Fähigkeitsniveau. Ein „Infit“

von 1 indiziert eine ideale Passung. In Anlehnung an das Vorgehen in PISA (Organisation für wirtschaftliche Zusammenarbeit und Entwicklung, 2012) wird ein WMNSQ von unter 0.80 bzw. von über 1.20 als starker Misfit gewertet. Ein WMNSQ von über 1.15 indiziert einen leichten Misfit (vgl. Pohl & Carstensen, 2012). Mit dieser Analyse wurde geprüft, ob die modellimplizierten Annahmen für jede der drei Schülergruppen zutreffen.

(3) *Differenzielle Itemfunktionalität.* Im dritten Schritt wurden DIF-Analysen (Holland & Thayer, 1988) durchgeführt. Dabei wurden die geschätzten Itemparameter für die drei Schülergruppen paarweise miteinander verglichen, wobei für mittlere Fähigkeitsunterschiede zwischen den jeweiligen Gruppen kontrolliert wurde. Es wurde geprüft, ob sich die Itemschwierigkeiten zwischen den betrachteten Schülergruppen unter Berücksichtigung von Fähigkeitsunterschieden bedeutsam voneinander unterscheiden. Das Ausmaß des DIF pro Item wurde wie folgt kategorisiert (vgl. Penfield & Camilli, 2007): geringer DIF wenn $|DIF| < 0.43$ oder nicht signifikant > 0; moderater DIF wenn $|DIF| \geq 0.43$ und $|DIF|$ signifikant > 0, sowie entweder $|DIF| < 0.64$ oder $|DIF|$ nicht signifikant > 0.43; starker DIF wenn $|DIF| \geq 0.64$ und signifikant > 0.43.

(4) *Vergleich der Korrelationsmuster zwischen Schülergruppen.* Um Hinweise auf die externe Validität der Testverfahren zu gewinnen, wurden im vierten Schritt korrelative Zusammenhänge der Testwerte in den Bereichen Lesen, Zuhören und Mathematik untereinander sowie mit verbalen und nonverbalen kognitiven Grundfähigkeiten zwischen den Schülergruppen verglichen. Um manifeste Indikatoren für die latente Personenfähigkeit zu gewinnen, wurden für jede Schülerin und jeden Schüler jeweils 15 *Plausible Values* pro Kompetenzbereich gezogen (von Davier, Gonzales & Mislevy, 2009). Zur leichteren Interpretierbarkeit wurden die Kompetenzwerte der Schülerinnen und Schüler auf die in LSAs übliche Berichtsmetrik ($M = 500$, $SD = 100$) transformiert. Die Analysen wurden für jeden der 15 *Plausible Values* pro Kompetenzbereich separat durchgeführt und anschließend nach Rubin (1987) zusammengefasst.

5. Ergebnisse

5.1 Forschungsfrage 1 – Anteil fehlender Werte

In Tabelle 2 sind die durchschnittlichen Anteile fehlender Werte an allen vorgelegten Testitems für die drei Schülergruppen in den Kompetenzbereichen Lesen, Zuhören und Mathematik abgebildet. Insgesamt unterscheidet sich der Anteil fehlender Werte zwischen den Gruppen erheblich. Bei Kindern ohne SPF war in allen untersuchten Bereichen der Anteil fehlender Werte durchgängig am kleinsten; bei Kindern mit SPF in Grundschulen war der Anteil jeweils ungefähr doppelt so hoch (Tab. 2). Schülerinnen und Schüler in Förderschulen hatten nahezu doppelt so viele fehlende Werte wie Kinder mit SPF in Grundschulen (Tab. 2). Insbesondere die Anteile fehlender Werte, die durch

	Deutsch Lesen				Deutsch Zuhören				Mathematik			
	Missing by Intention	*Missing Invalid Response*	*Missing not Reached*	*Missings insgesamt*	*Missing by Intention*	*Missing Invalid Response*	*Missing not Reached*	*Missings insgesamt*	*Missing by Intention*	*Missing Invalid Response*	*Missing not Reached*	*Missings insgesamt*
Kinder ohne SPF[1]	4.25%	0.79%	0.98%	6.02%	4.99%	0.99%	0.66%	6.65%	6.85%	1.09%	2.06%	9.99%
Kinder mit SPF in Grundschulen	9.05%	1.49%	2.28%	12.82%	8.47%	1.77%	1.97%	12.21%	11.84%	1.71%	4.58%	18.14%
Kinder mit SPF in Förderschulen	15.15%	2.32%	10.86%	28.33%	15.39%	2.19%	2.64%	20.22%	19.17%	1.74%	12.27%	33.18%

[1] sonderpädagogischer Förderbedarf

Tab. 2: Relativer Anteil fehlender Werte pro Fach und Kompetenzbereich in den drei untersuchten Schülergruppen

Auslassen *(Missing by Intention)* und durch Nichterreichen des Endes eines Testblocks *(Missing not Reached)* zustande kamen, waren in Mathematik und im Lesen bei Kindern mit SPF in Förderschulen deutlich höher als in den beiden anderen Schülergruppen (Tab. 2). Somit schienen insbesondere Schülerinnen und Schüler mit SPF in Förderschulen durch die Länge der Kompetenztests im Lesen und in Mathematik überfordert zu sein.

5.2 Forschungsfrage 2 – Passung des Raschmodells

Die Ergebnisse der Analyse zur Passung des Raschmodells sind in Tabelle 3 dargestellt. Für Kinder ohne SPF ergab sich für nahezu alle Items eine zufriedenstellende Passung (Tab. 3). Lediglich im Kompetenzbereich Mathematik wiesen einige Items (2 %) einen leichten Misfit auf. In der Gruppe der Kinder mit SPF in Grundschulen war der Anteil der Items mit einem Misfit vor allem in Mathematik höher (insgesamt 14 %). In der Schülergruppe mit SPF in Förderschulen schließlich waren anteilig noch einmal deutlich mehr Mathematikitems mit einem Misfit zu verzeichnen als in den anderen Schülergruppen (insgesamt 25 %). In den Kompetenzbereichen Lesen und Zuhören waren die Anteile der Items mit Misfit bei Kindern mit SPF in Förder- und Grundschulen etwas höher als bei Kindern ohne SPF (Tab. 3). Im Vergleich zum Kompetenztest in Mathematik lagen diese Anteile jedoch in einem akzeptablen Bereich (Tab. 3).

Außerdem wurde geprüft, ob die eingesetzten Testverfahren in allen Schülergruppen vergleichbar reliabel messen. Dies scheint der Fall zu sein: Die *Expected a posteriori*-Reliabilitäten (EAP-Reliabilitäten) als Maß für die durch das Raschmodell erklärte Varianz der Personenfähigkeiten im Verhältnis zur Gesamtvarianz der Personenfähigkeiten unterschieden sich in keinem der drei Kompetenzbereiche bedeutsam zwischen den Schülergruppen (Tab. 3).

5.3 Forschungsfrage 3 – Differenzielle Itemfunktionalität

Im Folgenden werden die Ergebnisse zur differenziellen Itemfunktionalität getrennt für die paarweisen Gruppenvergleiche zwischen Kindern mit SPF in Grundschulen, Kindern mit SPF in Förderschulen und Kindern ohne SPF dargestellt (Tab. 4).

Kinder mit SPF in Grundschulen vs. Kinder ohne SPF
Schülerinnen und Schüler mit SPF in Grundschulen erhielten dieselben Testhefte wie Schülerinnen und Schüler ohne SPF. Daher konnte in allen Kompetenzbereichen der gesamte Itempool auf DIF überprüft werden. Die Ergebnisse sind in Tabelle 4 zusammengefasst. Für alle drei Kompetenzbereiche zeigte sich erwartungsgemäß, dass die Items von Kindern mit SPF deutlich weniger häufig gelöst wurden als von Kindern ohne SPF (s. Linking-Konstante in Tab. 4). Im Lesetest wies keines der 80 Items einen starken DIF auf; bei 18 Items (23 %) wurde ein moderater DIF festgestellt. Im Kompetenzbe-

	Deutsch Lesen			Deutsch Zuhören			Mathematik		
	Kinder ohne SPF[1]	Kinder mit SPF in GS[2]	Kinder mit SPF in FS[3]	Kinder ohne SPF	Kinder mit SPF in GS	Kinder mit SPF in FS	Kinder ohne SPF	Kinder mit SPF in GS	Kinder mit SPF in FS
N Items: WMNSQ[4] < 0.80	1	0	0	0	0	0	0	10	4
N Items: WMNSQ > 1.15	0	1	1	0	1	1	6	9	3
N Items: WMNSQ > 1.20	0	4	1	0	0	0	0	19	4
N Items insgesamt	80	80	26	51	51	21	275	275	44
EAP-Reliabilität	.71	.73	.75	.65	.68	.70	.91	.92	.89

[1] sonderpädagogischer Förderbedarf, [2] Grundschulen, [3] Förderschulen, [4] standardisierter Weighted Mean Square (Wright & Masters, 1982); Kriterien zur Beurteilung des Itemfits: starker Misfit: WMNSQ < 0.80 oder WMNSQ > 1.20, leichter Misfit: WMNSQ > 1.15 und WMNSQ < 1.20 (OECD, 2012; Pohl & Carstensen, 2012)

Tab. 3: Itemfit pro Fach und Kompetenzbereich in den drei untersuchten Schülergruppen

	Deutsch Lesen			Deutsch Zuhören			Mathematik		
	Kinder mit SPF[1] in GS[2] vs. Kinder ohne SPF	**Kinder in FS[3] vs. Kinder ohne SPF**	**Kinder mit SPF in FS vs. Kinder mit SPF in GS**	**Kinder mit SPF in GS vs. Kinder ohne SPF**	**Kinder in FS vs. Kinder ohne SPF**	**Kinder mit SPF in FS vs. Kinder mit SPF in GS**	**Kinder mit SPF in GS vs. Kinder ohne SPF**	**Kinder in FS vs. Kinder ohne SPF**	**Kinder mit SPF in FS vs. Kinder mit SPF in GS**
Linking-Konstante (Logit)	1.03	2.14	1.14	0.76	1.92	1.10	1.05	2.19	1.08
Linking-Konstante (BISTA[4])	95.84	199.61	106.31	84.19	213.63	122.51	96.11	199.77	98.78
N Items mit sign. DIF[5]	19 (23.75 %)	8 (30.77 %)	5 (19.23 %)	16 (31.37 %)	8 (38.10 %)	5 (23.81 %)	60 (21.82 %)	26 (59.09 %)	10 (22.73 %)
N Items mit geringem DIF	62 (77.50 %)	20 (76.92 %)	19 (73.08 %)	41 (80.39 %)	14 (66.67 %)	13 (61.90 %)	220 (80.00 %)	25 (56.82 %)	33 (75 %)
N Items mit moderatem DIF	18 (22.50 %)	5 (19.23)	7 (26.92 %)	10 (19.61 %)	6 (28.57 %)	8 (38.10 %)	54 (19.63 %)	17 (38.64 %)	10 (22.73 %)
N Items mit starkem DIF	0 (0 %)	1 (3.85 %)	0 (0 %)	0 (0 %)	1 (4.76 %)	0 (0 %)	1 (0.36 %)	2 (4.54 %)	1 (2.27 %)
N Items insgesamt	80	26	26	51	21	21	275	44	44

[1] sonderpädagogischer Förderbedarf, [2] Grundschulen, [3] Förderschulen, [4] Bildungsstandards-Metrik, [5] Differential Item Functioning mit folgenden Kriterien zur Beurteilung des Ausmaß an DIF: geringer DIF: |DIF| < 0.43 oder nicht signifikant > 0, moderater DIF: |DIF| ≥ 0.43 und |DIF| signifikant > 0, sowie entweder |DIF| < 0.64 oder |DIF| nicht signifikant > 0.43; starker DIF: |DIF| ≥ 0.64 und signifikant > 0.43 (Penfield & Camilli, 2007)

Tab. 4: Differenzielle Itemfunktionalität pro Fach und Kompetenzbereich in den drei untersuchten Schülergruppen

reich Zuhören wies keines der 51 Items einen starken DIF auf; bei 10 Items (20 %) lag ein moderater DIF vor. Ein ähnliches Muster ergab sich für den Kompetenzbereich Mathematik, in dem nur ein Item mit starkem und 54 Items (20 %) mit moderatem DIF zu verzeichnen waren.

Kinder mit SPF in Förderschulen vs. Kinder ohne SPF
Im Gegensatz zu Kindern mit SPF in Grundschulen erhielten Schülerinnen und Schüler in Förderschulen verkürzte Testhefte. Eine vergleichende Analyse der Itemschwierigkeiten kann somit nur für diejenigen Items vorgenommen werden, die beiden Schülergruppen vorgelegt wurden. Im Vergleich zu Kindern ohne SPF waren die Items für Kinder in Förderschulen über alle Kompetenzbereiche hin deutlich schwieriger (Tab. 4). Der Anteil an Items mit einem starken DIF war zwar höher als im Vergleich zwischen Kindern mit SPF in Grundschulen und Kindern ohne SPF, die Anzahl der Testaufgaben, die dies betraf, war aber ebenfalls gering.

Kinder mit SPF in Förderschulen vs. Kinder mit SPF in Grundschulen
Auch bei den DIF-Analysen zum Vergleich von Kindern mit SPF in Förder- vs. Grundschulen konnte nur auf die Teilmenge des Itempools zurückgegriffen werden, die beide Schülergruppen bearbeitet hatten. Für Kinder mit SPF in Grundschulen waren die Aufgaben in allen Kompetenzbereichen leichter zu lösen als für Kinder in Förderschulen (Tab. 4). Die Differenzen in den Itemschwierigkeiten waren im Vergleich dieser Gruppen ähnlich groß wie die Unterschiede zwischen Kindern ohne SPF und Kindern mit SPF in Grundschulen (Tab. 4). Über alle Kompetenzbereiche hinweg wies jedoch lediglich ein Mathematikitem einen starken DIF auf. Der Anteil an Items mit moderatem DIF war dagegen etwas höher. Im Lesen zeigten sieben von 26 Items (27 %), im Zuhören acht von 21 Items (38 %) und in Mathematik 10 von 44 Items (23 %) einen moderaten DIF. Der geringe Anteil von Items mit einem starken DIF weist darauf hin, dass in den drei Kompetenzbereichen für Kinder mit SPF in Förder- und Grundschulen nach Kontrolle der Personenfähigkeiten keine substanziellen Unterschiede in den Itemschwierigkeiten auftraten, die Items demnach keine Gruppe systematisch „benachteiligten".

5.4 Forschungsfrage 4 – Vergleich der Korrelationsmuster zwischen den Schülergruppen

Im Hinblick auf die externe Validität der Testverfahren wurde geprüft, ob die Zusammenhänge der Leistungen im Lesen, Zuhören und in Mathematik untereinander sowie zu den verbalen und nonverbalen kognitiven Grundfähigkeiten zwischen den Schülergruppen ähnlich ausgeprägt sind. Tabelle 5 stellt die Ergebnisse der Korrelationsanalysen dar. In allen drei Gruppen korrelierten die Testwerte im Lesen, Zuhören und in Mathematik stark positiv miteinander. Allerdings waren die Korrelationen zwischen den sprachlichen Kompetenzbereichen und Mathematik bei Kindern in Förderschulen höher als in den anderen Schülergruppen (Tab. 5). Die Testwerte im Lesen und Zuhören

	1. Lesen[1]			2. Zuhören			3. Mathematik		
	Kinder ohne SPF[2]	Kinder mit SPF in GS[3]	Kinder mit SPF in FS[4]	Kinder ohne SPF	Kinder mit SPF in GS	Kinder mit SPF in FS	Kinder ohne SPF	Kinder mit SPF in GS	Kinder mit SPF in FS
1. Lesen									
2. Zuhören	.72	.71	.69						
3. Mathematik	.66[a]	.65[b]	.72[a, b]	.62[a]	.61[b]	.68[a, b]			
4. KFT[5] Verbal	.68	.71	.68	.68[a]	.70	.74[a]	.61[a]	.63	.69[a]
5. KFT Nonverbal	.46	.41	.41	.45	.41	.47	.53	.49	.54

[1] Testleistungen (erfasst mit je 15 Plausible Values), [2] sonderpädagogischer Förderbedarf, [3] Grundschulen, [4] Förderschulen, [5] Kognitiver Fähigkeitstest

[a] signifikanter Unterschied ($p < .05$, zweiseitige Testung) der Korrelation zwischen Kindern ohne SPF und Kindern mit SPF in Förderschulen

[b] signifikanter Unterschied ($p < .05$, zweiseitige Testung) in der Korrelation zwischen Kindern mit SPF in Förderschulen und Kindern mit SPF in Grundschulen

Tab. 5: Korrelationen zwischen schulischen Kompetenzen und kognitiven Grundfähigkeiten in den drei untersuchten Schülergruppen

hingen in den drei Schülergruppen stärker mit den verbalen kognitiven Grundfähigkeiten zusammen als mit den nonverbalen kognitiven Grundfähigkeiten (Tab. 5). Auch in Mathematik korrelierten die Testwerte stärker mit den verbalen als mit den nonverbalen kognitiven Grundfähigkeiten, jedoch waren die Zusammenhänge mit den nonverbalen kognitiven Grundfähigkeiten in Mathematik enger als mit den beiden Kompetenzbereichen im Fach Deutsch.

Mit paarweisen Vergleichen wurde geprüft, ob sich die Korrelationen zwischen den Gruppen signifikant voneinander unterscheiden (Tab. 5). Bei sechs von 27 Vergleichen fanden sich statistisch bedeutsame Unterschiede auf einem Signifikanzniveau von $\alpha = .05$. Diese betreffen in allen Fällen die Schülergruppe mit SPF in Förderschulen: Bei vier Vergleichen zeigten sich höhere Korrelationen für Kinder mit SPF in Förderschulen als für Kinder ohne SPF; bei zwei Vergleichen waren die Korrelationen für diese Schülergruppe höher als für Kinder mit SPF in Grundschulen (Tab. 5).

6. Diskussion

6.1 Zusammenfassung und Interpretation

Die Ergebnisse der vorliegenden Analysen weisen darauf hin, dass eine reliable und valide Testung von Schülerinnen und Schülern mit SPF in LSAs im Prinzip möglich ist. Einzelne Unterschiede in der Testgüte waren in der untersuchten Studie in erster Linie zwischen Kindern an Förderschulen und Kindern ohne SPF an Grundschulen zu verzeichnen.

Kinder mit SPF in Grundschulen

Für die Schülergruppe mit SPF in Grundschulen deuten die vorliegenden Ergebnisse darauf hin, dass sich die schulischen Kompetenzen auf einer gemeinsamen Skala mit den Testwerten der Schülergruppe ohne SPF abbilden lassen. Diese Kinder erhielten dieselben Aufgaben wie Schülerinnen und Schüler ohne SPF und bearbeiteten die Kompetenztests ohne Akkommodationen. Bei Anwendung des Raschmodells konnte eine akzeptable Passung der eingesetzten Aufgaben für Schülerinnen und Schüler mit SPF an Grundschulen in allen drei Kompetenzbereichen festgestellt werden.

Allerdings war der Anteil fehlender Werte in den drei Kompetenzbereichen deutlich höher als für Kinder ohne SPF. Diese Ergebnisse weisen darauf hin, dass die eingesetzten Tests für Schülerinnen und Schüler mit SPF zu umfangreich sind, um sie in der vorgegebenen Zeit vollständig beantworten zu können. Die Vergleichbarkeit der Testwerte durch differenzielle Anteile fehlender Werte wäre vor allem dann eingeschränkt, wenn diese mit konstrukt-irrelevanten Merkmalen zusammenhängen.

Die Ergebnisse der DIF-Analysen zeigen, dass Kinder mit SPF in Grundschulen durch die eingesetzten Aufgaben kaum eine Messbenachteiligung erfahren haben. Zudem ergaben Korrelationsanalysen für diese Gruppe ähnliche Zusammenhangsmuster der Testwerte im Leseverständnis, Zuhören und in Mathematik wie für Schülerinnen und Schüler ohne SPF. Ebenso korrelierten die kognitiven Grundfähigkeiten in vergleichbarer Stärke mit den Leistungen in den Kompetenztests. Diese Ergebnisse sprechen dafür, dass Testwerte von Schülerinnen und Schülern mit SPF in Grundschulen und Kindern ohne SPF in gleicher Weise das Kompetenzniveau reflektieren, das den Tests zugrunde liegt.

Kinder mit SPF in Förderschulen

Für die Schülerinnen und Schüler mit SPF in Förderschulen, die die Testverfahren mit den eingesetzten Akkommodationen bearbeiteten, scheint die Vergleichbarkeit der Messungen deutlich stärker eingeschränkt zu sein. Für die Kompetenzbereiche Lesen und Zuhören lässt sich anhand der Ergebnisse aus den Itemfit-Analysen schlussfolgern, dass das Raschmodell geeignet ist, auch die Antworten der Kinder mit SPF in Förderschulen auf die Items angemessen abzubilden. Im Kompetenzbereich Mathematik wies hingegen ein beträchtlicher Anteil der Items in dieser Gruppe einen Misfit auf. Die Abbildung der mathematischen Kompetenzen der Schülerinnen und Schüler an Förderschulen auf einer gemeinsamen Skala mit Schülerinnen und Schülern ohne SPF ist daher nur eingeschränkt möglich.

Erwartungsgemäß zeigte sich in allen betrachteten Kompetenzbereichen, dass die Aufgaben für Schülerinnen und Schüler mit SPF in Förderschulen deutlich schwieriger waren als für Kinder ohne SPF. Die Schwierigkeitsunterschiede zwischen Schülerinnen und Schülern mit SPF in Förder- und Grundschulen waren dagegen geringer ausgeprägt. Werden Fähigkeitsunterschiede zwischen den Schülergruppen berücksichtigt, legen die Ergebnisse der DIF-Analysen nahe, dass die Gütekriterien für die Population der Kinder mit SPF in Förderschulen durch die eingesetzten Aufgaben nicht systematisch eingeschränkt sind. Insbesondere im Vergleich der schulischen Kompetenzen von

Kindern mit SPF in Förder- und Grundschulen war der Anteil an Items mit einem starken DIF vernachlässigbar. Im Hinblick auf Aspekte der externen Validität der eingesetzten Kompetenztests zeigten sich für Kinder mit SPF in Förderschulen leicht abweichende Zusammenhänge der Testwerte im Leseverständnis, Zuhören und Mathematik im Vergleich zu Schülerinnen und Schülern mit SPF in Grundschulen bzw. der Schülergruppe ohne SPF. Die Testwerte in Mathematik korrelierten in dieser Schülergruppe stärker mit den Kompetenzen im Lesen und Zuhören als in den anderen Gruppen. Diese Ergebnisse schränken die Interpretation der Testwerte von Kindern in Förderschulen im Vergleich zu Schülerinnen und Schülern mit SPF in Grundschulen und Kindern ohne SPF möglicherweise ein.

6.2 Ausblick

Die von Schülerinnen und Schülern mit SPF in LSAs erreichten Leistungen werden durch deren Kompetenzen und individuelle Lernvoraussetzungen, aber auch durch Merkmale der Testaufgaben beeinflusst (Lane & Leventhal, 2015). Dieses Zusammenspiel zwischen Schüler- und Testmerkmalen sollte in zukünftigen Studien stärker berücksichtigt werden. Zum einen ist die Heterogenität der Schülerschaft mit SPF auch innerhalb der Schularten in Betracht zu ziehen. In den vorliegenden Analysen konnte beispielsweise keine differenzierte Betrachtung nach den verschiedenen Förderschwerpunkten vorgenommen werden, da die Teilstichproben zu klein waren. Eine entsprechende Analyse könnte Hinweise darauf geben, ob für Schülerinnen und Schüler in Abhängigkeit vom Förderschwerpunkt unterschiedliche Akkommodationen bereitgestellt werden sollten (Elliott et al., 2011). Außerdem dürften zwischen den Förderschwerpunkten Unterschiede dahingehend bestehen, ob die Schülerinnen und Schüler zielgleich oder zieldifferent unterrichtet werden (Autorengruppe Bildungsberichterstattung, 2014). Zieldifferent unterrichtete Schülerinnen und Schüler mit SPF haben spezielle Lehrpläne und werden daher mit den Anforderungen der an den Bildungsstandards ausgerichteten Kompetenztests weniger vertraut sein. Dies könnte die Vergleichbarkeit der Testwerte zwischen zielgleich und zieldifferent unterrichteten Schülerinnen und Schülern einschränken (Elliott et al., 2011).

Zum anderen sollte in weiterführenden Analysen geprüft werden, welche Aufgabenmerkmale (z. B. sprachliche Komplexität, Antwortformat) mit differenzieller Itemfunktionalität für Heranwachsende mit SPF zusammenhängen (vgl. Haag, Heppt, Stanat, Kuhl & Pant, 2013). Damit könnten potenzielle konstrukt-fremde Merkmale der Testaufgaben identifiziert werden, die bei der Aufgabenentwicklung berücksichtigt werden sollten.

Die vorliegende Untersuchung deutet darauf hin, dass der Einbezug von Schülerinnen und Schülern mit SPF in LSAs im Bildungsbereich prinzipiell möglich, teilweise jedoch mit Herausforderungen für die Testgestaltung verbunden ist. Die schulischen Kompetenztests müssen auf der einen Seite die Fähigkeiten der gesamten Schülerpopulation reliabel und vergleichbar abbilden. Auf der anderen Seite sollten die Leis-

tungsstände auch von Schülerinnen und Schülern mit SPF so akkurat und differenziert gemessen werden, dass die Ergebnisse auch für die Evaluation von Inklusionsbemühungen herangezogen werden können.

Die Generalisierbarkeit der Interpretation der Testwerte wird in der vorliegenden Untersuchung dadurch eingeschränkt, dass die eingesetzten Aufgaben für Kinder mit SPF im Durchschnitt deutlich schwieriger waren als für Kinder ohne SPF. Dadurch können Fähigkeitsunterschiede innerhalb der Schülergruppe mit SPF in Förderschulen im unteren Fähigkeitsbereich nicht mehr ausreichend differenziert werden (Südkamp, Pohl, Hardt, et al., 2015). Die vorliegenden Ergebnisse zeigen, dass Testverfahren, die nicht gezielt für diese Schülergruppe entwickelt und erprobt wurden, insbesondere im Fach Mathematik in Förderschulen zu schwer sind, um auch innerhalb dieser Gruppe eine reliable Interpretation der Testwerte zu ermöglichen. Dieser Herausforderung könnte durch die gezielte Entwicklung und Erprobung besonders leichter, auf die Gruppe der Schülerinnen und Schüler mit SPF abgestimmter, Testaufgaben begegnet werden.

Literatur

Abedi, J., Leon, S., Kao, J., Bayley, R., Ewers, N., Herman, J., & Mundhenk, K. (2011). *Accessible Reading Assessments for Students With Disabilities: The role of cognitive, grammatical, lexical, and textual/visual features. CRESST Report 785*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Autorengruppe Bildungsberichterstattung (2014). *Bildung in Deutschland 2014. Ein indikatorengestützter Bericht mit einer Analyse zur Bildung von Menschen mit Behinderungen*. Bielefeld: W. Bertelsmann.

Beauftragter der Bundesregierung für die Belange behinderter Menschen (Hrsg.) (2010). *Die UN-Behindertenrechtskonvention. Übereinkommen der Vereinten Nationen über die Rechte von Menschen mit Behinderung. Convention of the United Nations on the rights of persons with disabilities – deutsch, deutsch Schattenübersetzung, englisch. Bonn 2010.* http://www.behindertenbeauftragter.de/SharedDocs/Publikationen/DE/Broschuere_UNKonvention_KK.pdf?__blob=publicationFile [31. 10. 2014].

Böhme, K., & Robitzsch, A. (2009). Methodische Aspekte der Erfassung der Lesekompetenz. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 250–289). Weinheim: Beltz.

Bos, W., Müller, S., & Stubbe, T. C. (2010). Abgehängte Bildungsinstitutionen: Hauptschulen und Förderschulen. In G. Quenzel & K. Hurrelmann (Hrsg.), *Bildungsverlierer. Neue Ungleichheiten* (S. 375–397). Wiesbaden: VS Verlag für Sozialwissenschaften.

Chudowsky, N., & Pellegrino, J. W. (2003). Large-Scale Assessments That Support Learning: What will it take? *Theory into Practice, 42*(1), 75–83.

Cole, N. S., & Zieky, M. J. (2001). The New Faces of Fairness. *Journal of Educational Measurement, 38*(4), 369–382.

Elliott, S. N., Beddow, P. A., Kurz, A., & Kettler, R. J. (2011). Creating Access to Instruction and Tests of Achievement: Challenges and solutions. In S. N. Elliott, R. J. Kettler, P. A. Beddow & A. Kurz (Eds.), *Handbook of Accessible Achievement Tests for All Students* (pp. 1–16). New York: Springer.

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

Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H.A. (2013). Second Language Learners' Performance in Mathematics: Disentangling the effects of academic language features. *Learning and Instruction, 28*, 24–34.

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

Heydrich, J., Weinert, S., Nusser, L., Artelt, C., & Carstensen, C.H. (2013). Including Students With Special Educational Needs Into Large-Scale Assessments of Competencies: Challenges and approaches within the German National Educational Panel Study (NEPS). *Journal for Educational Research Online, 5*(2), 217–240.

Holland, P.W., & Thayer, D.T. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. In H. Wainer & H.I. Braun (Eds.), *Test Validity* (pp. 129–145). New York: Erlbaum.

Kocaj, A., Kuhl, P., Kroth, A.J., Pant, H.A., & Stanat, P. (2014). Wo lernen Kinder mit sonderpädagogischem Förderbedarf besser? Ein Vergleich schulischer Kompetenzen zwischen Regel- und Förderschulen in der Primarstufe. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 66*(2), 165–191.

Kornmann, R. (2006). Die Überrepräsentation ausländischer Kinder und Jugendlicher in Sonderschulen mit dem Schwerpunkt Lernen. In G. Auernheimer (Hrsg.), *Schieflagen im Bildungssystem. Die Benachteiligung der Migrantenkinder* (S. 71–85). Wiesbaden: VS Verlag für Sozialwissenschaften.

Lane, S., & Leventhal, B. (2015). Psychometric Challenges in Assessing English Language Learners and Students with Disabilities. *Review of Research in Education, 39*(1), 165–214.

Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8.

Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD). (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing.

Olson, J.F., & Goldstein, A.A. (1997). *The Inclusion of Students With Disabilities and Limited English Proficient Students in Large-Scale Assessments: A summary of recent progress*. Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement.

Opp, G., Budnik, I., & Fingerle, M. (2008). Sonderschulen – integrative Beschulung. In W. Helsper & J. Böhme (Hrsg.), *Handbuch der Schulforschung* (S. 341–361). Wiesbaden: VS Verlag für Sozialwissenschaften.

Penfield, R.D., & Camilli, G. (2007). Differential Item Functioning and Item Bias. In C.R. Rao & S. Sinharay (Eds.), *Handbook of Statistics*. Psychometrics (vol. 26, pp. 125–167). Amsterdam: Elsevier/North Holland.

Pitoniak, M.J., & Royer, J.M. (2001). Testing Accommodations for Examinees With Disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research, 71*(1), 53–104.

Pohl, S., & Carstensen, C.H. (2012). *NEPS Technical Report – Scaling the data of the competence tests*. Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Ricking, H. (2005). Der „Overlap" von Lern- und Verhaltensstörungen. *Sonderpädagogik, 35*(4), 235–248.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Smith, E., & Douglas, G. (2014). Special Educational Needs, Disability and School Accountability: An international perspective. *International Journal of Inclusive Education, 18*(5), 443–458.

Stanat, P., Pant, H.A., Böhme, K., & Richter, D. (Hrsg.) (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik: Ergebnisse des IQB-Ländervergleichs 2011*. Münster: Waxmann.

Südkamp, A., Pohl, S., Hardt, K., Jordan, A.-K., & Duchhardt, C. (2015). Kompetenzmessung in den Bereichen Lesen und Mathematik bei Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf. In P. Kuhl, P. Stanat, B. Lütje-Klose, C. Gresch, H.A. Pant & M. Prenzel (Hrsg.), *Inklusion von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen* (S. 243–272). Wiesbaden: Springer VS.

Südkamp, A., Pohl, S., & Weinert, S. (2015). Competence Assessment of Students With Special Educational Needs – Identification of appropriate testing accommodations. *Frontline Learning Research, 3*(2), 1–25.

Thurlow, M.L. (2002). Positive Educational Results For All Students: The promise of standards-based reform. *Remedial and Special Education, 23*(4), 195–202.

Thurlow, M.L. (2010). Steps Toward Creating Fully Accessible Reading Assessments. *Applied Measurement in Education, 23*(2), 121–131.

von Davier, M., Gonzalez, E., & Mislevy, R.J. (2009). What Are Plausible Values and Why Are They Useful? *IERI Monograph Series Volume, 2*, 9–36.

Winkelmann, H., & Robitzsch, A. (2009). Modelle mathematischer Kompetenzen: Empirische Befunde zur Dimensionalität. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 169–196). Weinheim: Beltz.

Wrase, M. (2015). Die Implementation des Rechts auf inklusive Schulbildung nach der UN-Behindertenrechtskonvention und ihre Evaluation aus rechtlicher Perspektive. In P. Kuhl, P. Stanat, B. Lütje-Klose, C. Gresch, H.A. Pant & M. Prenzel (Hrsg.), *Inklusion von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen* (S. 41–74). Wiesbaden: Springer VS.

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

Abstract: In 2011, students with special educational needs (SEN) were part of the sample of the National Assessment Study (IQB-Ländervergleich 2011) in primary schools. The present study examines if the achievement tests in German and Mathematics are also suited to assess SEN students' proficiencies. To evaluate the structural validity of the achievement tests for SEN students in special education and regular schools, missing patterns, the fit of the Rasch model, differential item functioning, and correlations of the test scores were analyzed. Results show that the achievement tests captured the same proficiencies in similar ways for SEN students compared to students without SEN. However, comparability of the achievement tests in Mathematics was reduced for students in special education schools.

Keywords: Special Educational Needs, Comparability, School Achievement, Primary School, Educational Placement

Anschrift der Autor_innen

Dipl.-Psych. Aleksander Kocaj, Humboldt-Universität zu Berlin,
Institut zur Qualitätsentwicklung im Bildungswesen,
Wissenschaftlicher Mitarbeiter und externer Fellow
der International Max Planck Research School on the Life Course (LIFE),
Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: a.kocaj@iqb.hu-berlin.de

Dr. Nicole Haag, Humboldt-Universität zu Berlin,
Institut zur Qualitätsentwicklung im Bildungswesen,
Wissenschaftliche Mitarbeiterin,
Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: nicole.haag@iqb.hu-berlin.de

Dr. Sebastian Weirich, Humboldt-Universität zu Berlin,
Institut zur Qualitätsentwicklung im Bildungswesen,
Wissenschaftlicher Mitarbeiter,
Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: sebastian.weirich@iqb.hu-berlin.de

Dr. Poldi Kuhl, Humboldt-Universität zu Berlin,
Institut zur Qualitätsentwicklung im Bildungswesen,
Leiterin des Forschungsdatenzentrums,
Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: poldi.kuhl@iqb.hu-berlin.de

Prof. Dr. Hans Anand Pant, Humboldt-Universität zu Berlin,
Institut für Erziehungswissenschaften,
Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: hansanand.pant@hu-berlin.de

Prof. Dr. Petra Stanat, Humboldt-Universität zu Berlin,
Institut zur Qualitätsentwicklung im Bildungswesen,
Direktorin und wissenschaftlicher Vorstand,
Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: iqboffice@iqb.hu-berlin.de